INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Previsão de anulação de projetos na gestão de fundos públicos

Afonso Chen Miao

Mestrado em Engenharia Informática

Orientador(a):
Doutor Luís Miguel Martins Nunes, Professor Associado,
ISCTE-IUL

Co-Orientador(a):
Doutora Elsa Alexandra Cabral Rocha Cardoso, Professora Auxiliar,
ISCTE-IUL

Outubro, 2023

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologias da Informação

Previsão de anulação de projetos na gestão de fundos públicos

Afonso Chen Miao

Mestrado em Engenharia Informática

Orientador(a):
Doutor Luís Miguel Martins Nunes, Professor Associado,
ISCTE-IUL

Co-Orientador(a):
Doutora Elsa Alexandra Cabral Rocha Cardoso, Professora Auxiliar,
ISCTE-IUL

Outubro, 2023

Acknowledgments

This research task was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020].

This research was only possible thanks to the support:

To my supervisor, Professor Luis Nunes, I thank him for all his guidance and dedication along the way. His commitment and dedication were crucial to my success. All his availability and advice were fundamental to the rigour of the dissertation.

To my co-supervisor, Professor Elsa Cardoso, I would like to express my appreciation for her guidance. Thank you for investing time and effort in my academic development.

To my parents and sister, I would like to express my deepest gratitude for their constant support. They have always been exemplary and have helped me to become the man I am today.

To my girlfriend for pushing me and not making me give up in difficult times. Thank you for all your words and affection throughout my academic career.

To my friends who motivated me to go out and relax at times when I felt most discouraged.

A sincere thank you to everyone I've mentioned.

Resumo

O recurso ao empréstimo de entidades terceiras para dar continuidade e desenvolvimento aos projetos é fundamental. No entanto, a aceitação das propostas apresentadas por múltiplas empresas requer uma análise detalhada, tendo especial atenção aos projetos de maior risco de cancelamento durante a sua vigência, devido às restituições monetárias. O estudo tem como objetivo prever com antecedência o risco associado a estes projetos, de modo a acompanhar com maior cuidado aqueles que aparentam ter um maior risco.

Ao obter o conjunto de dados final após a agregação de múltiplos ficheiros, foram aplicados métodos de seleção de *features* - *ANOVA F-value*, *Mutual Information*, *Feature Importance* - com maior impacto na classificação do projeto. Verificou-se também que havia um desequilíbrio entre as classes apresentadas no conjunto de dados e, portanto, foram realizados testes com o método *Synthetic Minority Over-sampling Technique* (*SMOTE*), para gerar dados sintéticos de forma a equilibrar o conjunto de dados.

Para obter o modelo com a melhor capacidade de classificar os projetos terminados dos anulados, foram realizados múltiplos testes com vários modelos, tipos de *features* e métodos de pré-processamento, totalizando 600 testes. No final, concluiu-se que o modelo *Multi-Layer Perceptron* com o método de pré-processamento *Standard Scaler*, com geração de dados sintéticos através do método *SMOTE,* redução da classe maioritária através de *random under sampling* e híper parametrização obteve o melhor resultado, obtendo um resultado de *f1* de 79%. Sendo assim, este modelo é considerado um modelo preditivo com potencial para ser utilizado neste tipo de problemas.

Palavras-Chave: Machine learning, Data mining, Anulação de projeto

Abstract

The use of loans from third parties to give continuity and development to companies' projects is fundamental. However, the acceptance of proposals submitted by multiple companies requires a detailed analysis, with particular attention to projects with a higher risk of cancellation during their term, due to monetary refunds. The study aims to predict the risk associated with these projects in advance, in order to monitor more carefully those that appear to have a greater risk.

Once the final dataset was obtained, after aggregating multiple files, methods were applied to select the features - ANOVA F-value, Mutual Information, Feature Importance - with the greatest impact on project classification. It was also found that there was an imbalance between the classes presented in the dataset and so tests were carried out using the Synthetic Minority Over-sampling Technique (SMOTE) method to generate synthetic data in order to balance the dataset.

To find the model with the best ability to classify completed projects from cancelled ones, multiple tests were carried out with various models, types of features and pre-processing methods. A total of 600 tests were performed. It was concluded that the *Multi-Layer Perceptron* model with the *standard scaler* pre-processing method, with generation of synthetic data using the SMOTE method, plus majority class reduction using random under sampling and hyper parameterisation obtained the best results, with 79% for f1 score. This model is therefore considered a predictive model with potential for use in this type of problem.

Keywords: Machine learning, Data mining, Project Cancellation

Index

Indexes of tables

Indexes of figures

Glossary of Abbreviations and Acronyms

ACC – Accuracy

AUC – Area under curve

CRISP-DM – Cross Industry Standard Process for Data Mining

DT – Decision Tree

EBITDA – Earnings before income tax and depreciation

ELM – Extreme Learning Machine

FIRR – Financial internal rate of return

FNPV – Financial net present value

GNN – Gaussian Naïve Bayes

IAPMEI – Instituto de Apoio às Pequenas e Médias Empresas e à Inovação

KNN – K-Nearest Neighbours

LR – Logistic Regression

LSVC – Linear Support Vector Classification

MLP – Multi-Layer Perceptron

ML – Machine learning

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

ROI – Return on investment

ROE – Return on equity

RF – Random Forest

SGB – Stochastic Gradient Boosting

SLP – Single-Layer Perceptron

SME – Small medium enterprises

SVC – Support Vector Classification

SMO – Sequential Minimal Optimization

XGBoost – Extreme Gradient Boosting

CHAPTER 1

Introduction

In order to increase the competitiveness of companies, non-repayable grants are available from institutions such as IAPMEI, I.P (the Agency for Competitiveness and Innovation[1]). This support is only guaranteed once the application has been analysed and accepted by the institution. Therefore, the application analysis phase is an extremely important stage in the application for support, in order to guarantee that the support will be given to a project with a good capacity for successful completion.

The study concentrates on this phase of a project application using the data provided by IAPMEI. The aim was to create an automatic system for analysing applications using machine learning predictive models.

Although the data provided was structured, it had to be processed and standardised in order to obtain the desired format for the work. The data was processed, and then new variables were calculated from the existing ones in order to improve the results obtained.

Following the processing of the data set, the experiments that were to be conducted were defined with the aim of determining the optimal machine learning model and its characteristics. Using particular techniques, the features that had the biggest influence on the outcomes were extracted.

Therefore, the last phase focused on running the experiments and analysing the results of the various models and features used. All these experiments were executed with python recurring to the library of scikit-learn to train the models, extract the most impactful features and listing the results of each tested model.

The models developed will make it possible to predict projects with a lower risk of being cancelled in a more complete, rapid and automatic way during the analysis and decision-making process for delivering project support to a company. As well as speeding up the evaluation process, it will significantly contribute to reducing project cancellations, thus optimising the allocation of resources and promoting more successful project results.

With the development of technology came the increase in data as a result of its extraction. Data has become an asset for organizations since insights can be gained from it. Therefore, data is an asset of great importance. The development of reliable models with good accuracy is only possible if large volumes of high-quality data is available. Data has therefore become very important for organisations because of the ability to obtain information from it, but it is also indispensable for decision-making.

[1] In Portuguese, Agência para a Competitividade e Inovação.

Analysing applications for public funds is still a manual process, lacking efficient tools to speed up analysis and improve decision-making. One of the crucial stages to improve is the evaluation and management of applications for public funds, where the lack of simplified mechanisms poses challenges to speeding up acceptance decisions and mitigating the risks associated with potential project cancellations.

The adoption of artificial intelligence technologies to assist the various tasks that occur in our daily lives is growing. The use of machine learning models has proven to be effective and above all a great aid tool for decision making [1]. In order for the results obtained by these models to be interesting and reliable, it is important to train and test the model created to validate its accuracy, analyse and extract the relevant data according to the problem to be solved. Only with appropriate and clean data can an effective and coherent machine learning model be achieved.

Over several years, the prediction of non-performing loans and the insolvency of a company given its current state were made through statistical data analysis techniques. The purpose of this study is to help public lender entities to verify if the third-party entity (or beneficiary) has all the characteristics and conditions to commit to what is initially agreed upon and required by the loan or grant applications. Adopting the use of machine learning models to analyse the loan or grant application could bring interesting and promising results for these financial institutions, being a complimentary tool for their decision to application acceptance. The main focus is on creating a strong machine learning model that can be used to forecast the likelihood of funding requests being cancelled. This offers funding agencies a chance to completely change the way they make decisions. It is important to stress that machine learning models do not rule the final decision, nor exclude the technicians' intervention, but may become an important tool for decision support in this task.

1.1. Objective and research questions

The objective of this study is to determine to what extent can a ML model predict if the project presented by an organization to a funding entity will be cancelled or completed. Additionally, it is intended to validate the financial ratios associated with companies that indicate signs of insolvency and eventually corruption. In order to achieve this, it will be necessary to train machine learning models with existing data (from an ongoing project), use multiples metrics to test and verify the precision of the model developed and apply them to new data that was not used when training the model to verify the results.

One model that can be used in the application analysis provided to the funding organization is a machine learning model that displays metrics with precise outcome indications for a collection of test data.

The generated model can be used as a decision-making tool when the debtor project promoter presents their expenses. In this way, it will be possible to anticipate, streamline and speed up analysis.

This following list contains all the questions that this study will address:

1. What conclusions and outcomes can be drawn from previous research that are relevant to the prediction of expense funding cancellation?
2. Are financial ratios used in business failure useful to predict the project cancelation?
3. Which information, besides financial ratios, have a strong impact on project cancellation?
4. What are the ratios of a company that have a strong link to project cancellation?
5. What is the performance of machine learning models on the prediction of project cancellation using the selected financial ratios?

CHAPTER 2

Literature review

One of the objectives of loans made by entities such as IAPMEI, I.P (the Agency for Competitiveness and Innovation[2]), public funds or banks is the acceleration, enterprise, and development of projects, and this is only possible if the borrower applies for these loans.

As is stated above, loans are a very impactful financial instrument for the growth and expansion of an enterprise and so funders (usually state surrogates) lend money to companies with the promise that the value will be repaid in the future. Grants are an additional source of funding besides loans, being a financial instrument, which differs from the typical loan in that repayment is not required. However, the loan can be beneficial depending on the type of funding the project needs. The grant does not cover all types of costs and is limiting in how it is used, while the loan does not have this limiting factor in how it is spent. Yet, these types of funding are time consuming and complicated to achieve as they require the submission of forms and proposal, which discriminates costs for the action, and consequently the analysis of project viability.

Before the approval of any credit or grant application public funds need to analyse the status of the company, the project and the risk associated with its implementation. As a result, applications from businesses that exhibit a precarious financial situation should be less likely to be approved. So, a crucial phase in the credit approval research is the investigation of the financial condition using financial ratios.

Over time, several studies have been conducted to estimate company bankruptcy, using data from financial ratios and statistical models to complement each other and make the prediction. The evolution of information systems and the growth of the volume of data currently available have helped in the development and adoption of machine learning models to predict while helping on analysis tasks. Thus, these types of models can also make prediction of an enterprise bankruptcy through multiple financial ratios.

The literature review is related to the studies mentioned above: how to analyse an enterprise health through financial ratios, impact of Machine Learning to predict business bankruptcy, techniques to extract the most correlated bankruptcy ratios, and variables influencing the cancellation of EU funded projects.

[2] In Portuguese, Agência para a Competitividade e Inovação.

2.1. Methodology

For this study, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses methodology) was applied for the literature review. The methodology consists of improving the systematic review given previously selected articles. The following search engines were used to perform the search for articles: Google Scholar, IEEEXplore and ACM Digital Library. All the searches for the articles were made in English, considering the articles most relevant from each search engine. No articles or studies directly related to the intended topic were found. Consequently, the literature review search was segmented into sub-themes related to the main objective. The following Table 1 refers to the keyword combinations used for each sub-theme:

Table 1. Words combination used to search for subthemes.

Subtheme	**Words combination**
How to analyse an enterprise health through financial ratios	Corporate Failure prediction + Financial Ratios Default prediction + Small enterprises Business failure + risk analysis
Types and characteristics of enterprises expenses	Cancelled projects + European Funds Eligible Costs + European Funds
Impact of Machine learning to predict business bankruptcy	Credit risk or Prediction of bankruptcy + machine learning Artificial Intelligence + finances

A total of 72 articles were collected from the aforementioned databases and selected according to search engine relevance, of which 6 were removed due to duplication. Within the 66 articles 40 were excluded as they did not meet the intended topics. In the end, 1 article was eliminated because the full text could not be found, and so, after reading the full text of the 25 articles, only 21 were selected, the reason for this exclusion being that they did not have useful information for the research questions addressed in this project. The 4 articles that were not selected are mentioned in Table 2.

Table 2. Articles removed from review

Author	**Article**
Florian MARIN	The Problems of Absorption of the European Structural and Investment Funds Related to the Cohesion Policy during the Programming Period 2014-2020, Review of International Comparative Management, 2019

Noor Hazlina Ahmad, Pi-Shen Seet	Dissecting Behaviours Associated with Business Failure: A Qualitative Study of SME Owners in Malaysia and Australi, September 2009
Emma H. Wood	The internal predictors of business performance in small firms, Journal of Small Business and Enterprise Development, 2006, Vol.13
Sung-Wook Kang	An Identification of Unsuccessful, Failure Factors of Technology Innovation and Development in SMEs: A Case Study of Components and Material Industry, International Journal of Business and Management, 2012, Vol.7

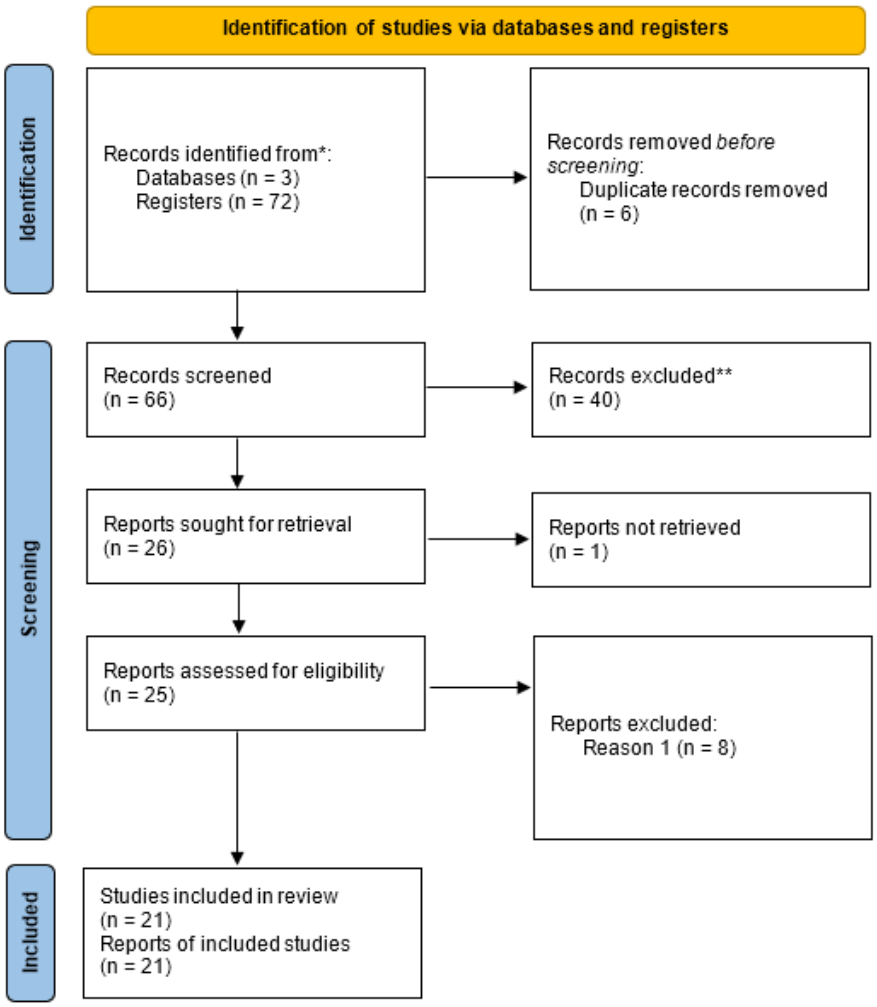

Figure 1. The PRISMA flow diagram.

2.2. How to analyse an enterprise health through financial ratios

Roy A.Foulke defined ratios as “a figure or a percentage representing the company of one dollar amount with some other dollar amount as a base.” [2]. To obtain meaningful information about an enterprise, financial ratios are constructed using numerical values collected from financial statements. Therefore, financial ratios have been shown to properly distinguish between failed and non-failed enterprises several years before failure, and they have long been regarded as reliable indicators of company failure [3]. Financial failure, or technical insolvency, is defined as the inability of a going concern to pay its debts as they are due [4].

Of the first techniques used to carry out forecasting was univariate analysis by Beaver. Using a series of financial ratios, he suggests a binary categorization model of businesses (bankrupt/non-bankrupt) [3]. The financial ratios considered good predictors of failure are cash flow/total assets, net income/total assets, total debt/total assets, and cash flow/total debt [2]. The latter ratio – cash flow /total debt – was identified as the strongest predictor ratio in terms of error rate, and it resulted in a classification accuracy of 78% five years prior to bankruptcy [3].

Succeeding univariate analysis method, Altman (1968) applied multiple discriminant analysis (MDA) for the prediction of corporate bankruptcy and were selected the following five ratios: working capital/total assets, retained earnings before interest and taxes/total assets, equity market value/book value of total debts and sales/total assets. Each ratio was classified into one of the following five ratios categories: liquidity, profitability, leverage, solvency, and activity ratios. On a sample of 66 holdout companies, Altman's method had a 79% success rate in predicting failure one year in advance [5].

According to Deakin's study, in terms of error classification rates, the Altman model performs better than the Beaver model within one year, but the Beaver model excels when the time frame is five years. Therefore, the latter model has a better accuracy in long-term analysis, while Altman's model excels in the short term. As mentioned above, the financial ratios used by these two models differ and according to Holmen the study of Beaver's univariate model using cash flow/total debt as ratio proved to have a lower error rate compared to the Altman model [3].

The results of the study indicate that all models are statistically significant one year prior to bankruptcy with the ratio model being the most effective in predicting the likelihood of bankruptcy. The presence of cash-flow ratio in a predictive model reinforces the warning of bankruptcy, while all models that did not present cash-flow ratio are found to be insignificant in years two and three before failure [6].

In the study of article [6] the most reliable indicators are working capital/total assets (liquidity ratio), total debt/total assets (leverage ratio), earnings before interest and tax/total assets (profitability ratio) and operating cash-flow/current liabilities (cash-flow ratio). The statistical method

used to design two models is logistic regression. The first model consists of the following ratios – activity, liquidity, leverage, profitability and cashflow. The second model was based on the first model but the ratio cashflow was removed. According on the results presented [6], [7], the model that contained cashflow ratio had a better overall accuracy (68.30%) compared to the other model (66.90%). Therefore, the ratios which led to a better prediction of bankruptcy are working capital/total assets (liquidity ratio), total debt/total assets (leverage ratio), earnings before interest and tax/total assets (profitability ratio) and operating cash-flow/current liabilities (cash-flow ratio). Hence, failure is correlated with a rise in leverage and a decline in cash flow, profitability, and liquidity.

In a previous study it was concluded that the weight of selected financial ratios has influence on the accuracy of the model according to the enterprise size. Applying discriminant analysis by size and business categories led to an increase of correctly classified firm's percentage [5]. The models developed for a specific industry presented better results in comparison to a generic model, which means that isolating by industry and size of the enterprise leads to an improved default prediction accuracy [8].

According to a study by [8], there are key indications of SME (small and medium enterprises) default relating to annual sales growth, operating revenue, and profitability growth.

Table 3 represents the models and financial ratios used and the results obtained in previous studies.

Table 3. Models and financial ratios used to predict bankruptcy.

Article	**Financial Ratios**	**Best Results (Accuracy)**	**Notes**
[6]	working capital/total assets (liquidity ratio) total debt/total assets (leverage ratio) earnings before interest and tax/total assets (profitability ratio) operating cash-flow/current liabilities (cash-flow ratio)	Logistic Regression (LR) – 68.30%	
[5]	Total Debts / Equity (Leverage) Bank Loans / Turnover (Leverage) ROE (Profitability) ROI (Profitability)	Logistic Regression (LR) - 88%	80% (Applied without distinction of category, size and group), 85 % (each size

	ACID TEST RATIO (ATR) (Liquidity)		group) and 88% (separate through manufacturing category)
[7]	Working capital to total assets (WCTA) Debt ratio (DR) Cash flow from operating activities to total assets (CFFOTA)	Logistic Regression (LR) - 77.86% (1 year before the failure)	72.14% (2 years before the failure) 74.29% (3 years before the failure)
[9]	Working capital/total assets Retained earning/total assets Earnings before interest and taxes/total assets Market value of equity/book value of total debt Sales/total assets	Z-Score Model – 43%	Score of 73% in type II error rate
[9]	Return on Investment (ROI) Capital Turnover (CT) Inventory Turnover (IT) Financial Leverage (FL) Receivables Turnover (RT) Short Term Liquidity (STL) and Cash Position (CP)	Probit Model – 75%	Prediction on business failure and turnaround
[10]	Retained earnings*/total assets* Interest and discount expense/ (Short term borrowings + long term borrowings + corporate bond +convertible bond + note receivable discounted) Note payable + accounts payable) * x12/Sale (Current period liabilities and shareholders equity/Previous	Multivariate discriminant analysis model – 86.14 %	Independent of industry and size

	period liability and shareholders equity)-1		

2.3. Types and characteristics of enterprises expenses

As is quoted in the guideline [11], the principles that are applied to grants are: co-financing principle, no double financing rule and no-profit rule. The first rule states that Community grants could not cover all the expenses associated with the action that will be subsidized and the candidate also needs to provide recursive or financial support. The second rule states that the applicant should only have one grant, it is not allowed to have a double European Community funding for the same expenditure. The last rule mentions that the grant application cannot be aimed to produce profit for the applicant.

When requesting funding from the EU, a cost-benefit analysis is necessary to ensure the efficient allocation of public funds. This kind of study uses a quantitative methodology to assess a project's opportunity and requirement using cost and future benefits calculations.

If the request is granted funding, the commission will determine the contribution as a percentage of all qualifying expenses as listed in projected budget. An applicant that cooperates with its financial support on project shows a higher incentive in achieving better results. So, in many instances the percentage of the eligible costs must not be more than 80%. Therefore, the remaining 20% is up to the applicant or organization to cover up through cash, own sources, sponsorships, or bank loan.

Finally, if the proposal is accepted by the European Community, the distribution of the grant is done in two stages – pre-financing payment and final payment. In the first stage, the applicant will receive 50% of the payment upon signature of the agreement. In the final stage, the process to determine the remaining and final grant amount relies on the submission of the final technical implementation report and final financial statement. Due to this, the cost associated with carrying out the activity must be supported by invoices or other comparable supporting papers to be regarded as real expenditure.

The combination of eligible and ineligible costs prefixes the total value of the project. For the costs to be considered eligible it needs to meet certain conditions as listed below:

- be spent during the action;
- must be specified in the expected overall budget of the grant agreement;
- be required to carry out the project;
- be recognizable and verifiable, in particular appearing in the beneficiary's accounting records;
- obey the rules set forth by the relevant tax and social laws;
- be justifiable, reasonable, and adhere to the standards of solid financial management.

The following expenses are regarded as eligible direct costs: staff, travel, accommodation and subsistence allowances, costs of services, subcontracting and administration costs. While costs as: VAT; contributions in kind; debt; return on capital; interest and other charges on loans; penalties and court costs; and any allowable costs that surpass set limits are considered ineligible and are therefore not covered by the grant.

2.4. Variables influencing the cancellation of EU funded projects

In [12] two types of models were tested to predict the cancellation of projects supported by the European fund – linear probability model (LMP) and probit model. The variables presented in the study – fyzos[3], y2008[4], avg_emp[5], industry[6], age[7], paid[8], cancelled[9] – did not present a correlation higher than 25%. The correlation between avg emp and paid is the greatest one: 23.07% being an acceptable low value.

The variables or determinants that had an impact on cancelled projects are paid, age, avg_emp and y2008. The variables that had a negative sign (decreasing the probability of cancellation) in both models were paid and age. Thus, the project being likely cancelled was lower the larger the paid allocation. The reason behind this behaviour could be "that projects with large sums are of much bigger interest to potential beneficiaries, therefore, the project managers may be more focused on the project finalization." The second variable – age – had a negative impact in the LPM and Probit models, and it concluded that an extra year of age reduces the likelihood of project cancellation by 0.53% and 0.5%, respectively [8], [12]. This supports the hypothesis that established businesses are probably more cautious when beginning a new project. The rest of the variables (avg_emp and y2008) had an increase on the probability of cancellation on both models, which implies that a company with larger staff and if it is a year in crisis the probability of projects cancellation significantly. Although, the significance level is higher on y2008 variable compared to avg_emp. The 2008 crisis, which corresponds with the global financial crisis from 2007 to 2009, had an impact on defaulted loans for SMEs and intensified the issues associated with credit obtention [8].

Hence, enterprises with more seniority but with smaller number of employees have more chance to finalize a project when applied for large grant [12].

[3] Denotes if a subject applying for a structural funding is a natural person
[4] Dummy variable that denotes the year of allocation (year when the economic crisis broke out)
[5] Average number of employees per firm
[6] Equal to one if the main firm's activity is in industry or in the primary sector, and zero if a firm operates in services (of various kinds)
[7] Number of years since the establishment of a firm
[8] Sum of money that was paid to the beneficiary
[9] Binary variable to represent if a project was cancelled or finalized

In addition to the presented variables of a company that influence the prediction of project cancellation, the calculation of the following indicators – FNPV and FIRR - also allow to conclude if a project it is efficient, feasible and lucrative. Hence, a positive FPNV and value of FIRR greater than the discount rate and subunit cost/benefit ratio, distinguishes a successfully completed project from a cancelled one [13].

In a survey made for 186 Wester Pomerania SME, 78% were constituted by enterprises which project were financed by UE budget, the top 3 answers to the question which mentions the risks that enterprises encounter during the implementation of a project were: miscalculations on costs leading to additional costs that were not in project budget (36%), ineligible expenditure (27%) and issues with the supplier or contractor's contact (21%). The first and second answers are related since excluding any project costs results in ineligible spending [14].

Besides all the ratios mentioned above, there are a few factors which improves the results in default prediction. Non-financial factors such as employees' loyalty/satisfaction and long-term relationships with customers, credit behavioural information, management skills, education and expertise of the SME can lead to a better default prediction model which could also be fundamental factors for the project's cancellation. Access to the payment history of small firm with a duration of more than one year also shows improvements in the forecasting results. However, the use of these quality variables alongside with financial ratios is limited when using traditional methodologies [8].

2.5. Impact of machine learning to predict business bankruptcy

Mathematically speaking, it is generally known that simpler statistical learning models – linear regression, logistic regression and naïve bayes models – have a high interpretability but may have a low prediction accuracy. As opposed to that, neural networks and tree models offer higher predictive accuracy but with limited interpretability.

An extreme gradient boost model is considered a widely used and fast machine learning model, getting a better predictive accuracy comparing to a more classic model, logistic regression. Using XGBoost model the predictive accuracy suffered an increase of 14.81% (0.81 to 0.93) [15].

The four factors that have the biggest bearing on the outlook for non-bankrupt companies are: profits before taxes plus interests paid, and earnings before income tax and depreciation (EBITDA), trade receivables and total assets. On the other hand, the factors that reveals more impact for defaulted companies are total assets; shareholder's funds plus noncurrent liabilities; profits before taxes plus interests paid. In brief, the probability of default increases when low total assets, limited shareholder funds and low proficiency are coupled. According to Shapley's composition the variable which presented a strong positive importance was total assets to total liabilities (the leverage),

proceeded by variables that are measures of operational efficiency – EBITDA and profit before taxes plus interest paid – and related to solvency – trade receivables [15].

In a test performed with statistical, machine learning and deep learning methods on German and Australian credit data, it showed that machine learning techniques are generally more accurate than statistical techniques, however it was concluded that deep learning methods were the most potent among them all. Based on metrics of AUC (Area under curve) and ACC (accuracy) to evaluate the performance of the models, the models that presented the best AUC on each dataset (German and Australian) were bagging (machine learning model) and rf (machine learning model – random forest); and the best ACC were ann (deep learning) and ELM, respectively. The machine learning models that are most used are AdaBoost, SVM, Tree-related, k-NN and Bagging, while SGB (Stochastic Gradient Boosting) and ELM (Extreme Learning Machine) have a relatively low citation. In the deep learning field, ANN and MLP are the models most used [16].

A basic empirical method of initial variable selection followed by a stepwise strategy to choose the variables for the final discriminant function was employed in several earlier research. The ability of this research [17] to offer general conclusions about the financial factors that can reliably forecast financial distress, however, is constrained. In order to battle that, the method ReliefF assigns significance to features based on their capacity to distinguish between comparable samples, which are determined by closeness in feature space. While irrelevant features keep weights that are almost zero, relevant traits acquire high positive weights. Thus, the influence of the characteristic on induction increases with the value of the ReliefF scores. Applying this method on a dataset of 150 failed and solvent Greek firms (2003-2004), the following characteristics significantly affect induction: WC/TA, EQ/CE and GRNI. The attributes applied in training set of the machine learning model are: WC/TA, EQ/CE, GRNI, SIZE, GRTA, TD/EQ, S/CE, COLPER, S/EQ, CE/NFA, PAYPER, INVTURN, and GIMAR.

In this dataset experiments with several learning models – naïve bayes, local decision stump (DS), RIPPER, RBF (neural network) and Sequential Minimal Optimization (SMO) - were applied, which led users to be able to anticipate bankruptcies with satisfactory accuracy within one year before the actual bankruptcy. The tests were carried out with several years - one, two and three - until the financial distress and the models that presented a marked decrease in the forecast from the second year onwards were Local DS, RIPPER and SMO. Users of learning algorithms were shown to be able to anticipate bankruptcies with satisfactory accuracy long before the actual bankruptcy with the use of learning algorithms. The model RIPPER presented the best prediction accuracy – 74% - in one year prior to financial distress, while naïve bayes obtained a better prediction accuracy – 68% - among them all in three years prior to financial bankruptcy.

The prediction model reveals that ratios as working capital to total assets (liquidity), equity to capital employed (capital structure) and net income growth (profitability growth) are crucial for

obtaining interesting results. Other factors that are included in the models but could impact positively would be qualitative variables as leadership, reputation, management's ability, and future prospects.

2.6. Techniques to extract the most correlated bankruptcy ratios

Several techniques are used to select the financial ratios that have the greatest impact for predicting the bankruptcy of a company – classification and regression tree model (CART), stepwise, *reliefF*, univariate analysis and *shapley*. The first two techniques were used separately in [10] and led to a similarity between the CART and Stepwise procedure's selection of variables. The ratios indicating the best discriminant power varied for each method. While the CART model selected retained earnings to total assets as the strongest predictor, Stepwise predicted Equity to Liabilities and shareholders.

The stepwise technique is limiting because it cannot provide generalized results [17]. Thus, it was used an alternative approach – *ReliefF*. This technique consists of assigning importance values to features based on their capacity to distinguish between comparable samples, which are determined by closeness in the feature space. Features that present a large *ReliefF* value have a greater impact on predicting enterprise bankruptcy. The financial ratios that show stronger influence on bankruptcy are WC/TA, EQ/CE and GRNI.

In [6] it was used a statistical analysis, univariate analysis. However, to avoid multicollinearity while choosing strong predictive financial ratios, the classification of ratios into different categories is the solution adopted. "The most predictive variables are working capital/total assets (liquidity ratio), total debt/total assets (leverage ratio), earnings before interest and tax/total assets (profitability ratio) and operating cash-flow/current liabilities (cash-flow ratio)."[6]

2.7. Summary of the systemic review

The financial ratios displayed in the Table 4 are included into the most pertinent articles and will serve as the foundation for developing features for the intended use. The calculation of financial ratios is done from documents such as balance sheet, income and cash flow statements. Financial ratios can be categorized into four types - Liquidity, Solvency, Profitability and Efficiency. Liquidity reveals a company's capacity to settle its current financial obligations. Values lower than one mean that the company have more liabilities than assets. Therefore, companies with a positive liquidity ratio should present at least a value of 1 [18]. Solvency ratios differ from liquidity with respect to the time frame of the payments of their obligations. Solvency reveals a company's capacity to fulfil its commitments over the long term. Therefore, solvency ratios give a broad overview of the company's indebtedness and indicate if cash flows are sufficient to pay interest charges and fixed costs such as rent and leases [19]. Profitability ratios show a company's capacity to produce profits as a return on investment made over a specific time period. Therefore, this financial ratio is typically used to compare a company's profitability from one period to another. The ratios show how well a business is using its resources to make a profit while also indicating the competitive situation of the company [18]. The latter, evaluates company's capacity to manage its assets and how effectively it used them to produce income [20]. Financial ratios related to this type are Accounts Receivables turnover, Creditors turnover and Inventory turnover.

The literature review has shown that there are financial ratios that only make sense for listed companies - shareholders' funds plus noncurrent liabilities / Fixed Assets, Shareholder's equity to capital employed, Total debt to shareholder's equity capital, Sales divided by shareholder's equity capital. The mentioned financial ratios do not make sense to be implemented in this project since the dataset is related to small and medium sized companies, i.e., companies that are not listed. Besides the exclusion of the mentioned financial ratios, the financial ratios "Gross income divided by sales" and "Creditors turnover" were adapted according to the information provided in the dataset.

Table 4 below shows the financial ratios which led to the best result on each article of corporate failure prediction:

Table 4. Best performing financial ratios.

Article	Financial Ratio	Formula	Description	Category
[21]	EBITDA	Net Income + Taxes + Interest Expense +	EBITDA is net income (profits) plus the reimbursable costs of interest, taxes, depreciation, and	Profitability

		Depreciation & Amortization	amortisation. Regardless of the depreciation assumptions or financing strategies utilised by the firms, EBITDA may be used to analyse and compare the underlying profitability of those companies [22].	
[21]	EBIT	Net Income + Interest + Taxes	EBIT determines a company's net income before income tax and interest expenditures are subtracted. EBIT is a tool for evaluating a business's core operations [22].	Profitability
[23]	Total Assets	Non-Current Assets + Current Assets		-
[23]	Total Liabilities	Non-Current Liabilities + Current Liabilities		
[24]	Total Assets to Total Liabilities	Total Assets = Liabilities + Owner's Equity) / (Total Liabilities)	A higher ratio indicates that the company has more assets relative to its liabilities, which suggests a stronger financial position and lower risk of insolvency	Coverage
[21]	Working capital	Total Current Assets - Total Current Liabilities	When a company's working capital is lower than one, it means that has more current liabilities than current assets. An enterprise that has sufficient working capital is able to finance both its ongoing operations and its expansion plans [22].	Liquidity
	Working capital divided by total assets	Working capital / Total assets	This ratio is a reliable indicator of company difficulties. In the absence of sufficient current assets, a company with negative working capital would likely have trouble	Liquidity

			satisfying its short-term obligations [25].	
[26]	Growth rate of net sales	[(current net sales - prior sales period net sales) / prior sales period net sales] x 100	How much did net sales grew compared to other years	Profitability
[26]	Growth rate of total assets	[(current total assets - prior sales period total assets) / prior sales period total assets] x 100	How much did total assets grew compared to other years	Efficiency
[21]	Accounts Receivables turnover	Net Sales / Average Accounts Receivable	The average number of times a firm collects its accounts receivable amount is measured by the accounts receivables turnover ratio. It is a measurement of how well a business manages its line of credit procedure and collects unpaid bills from customers. A company's accounts receivable turnover ratio is greater for an efficient business and lower for an inefficient one [22].	Turnover
[21]	Average collection period for receivables	(Accounts receivable / Net Credit sales) * 365 days		Liquidity
[27]	Capital employed	Total assets - Current Liabilities	The total amount of money used by a company or project to acquire profits is known as capital employed. The value of all the assets a firm uses to produce	-

			earnings is another way to define capital used. It is possible to determine how much has been invested by looking at the capital employed [22].	
[28]	Net fixed assets	Fixed assets after tax	An organisation seeks to employ fixed assets over the long term to help produce income. Intangibles are amortised, whereas fixed assets are subject to depreciation to account for the decrease in value as the assets are utilised. Therefore, net fixed assets is the asset's residual value of the fixed asset [22].	-
[27]	Capital employed to net fixed assets	Capital employed / net fixed assets		Efficiency
[29]	Creditors turnover	net credit / average accounts payable	A short-term liquidity metric that is used to quantify how quickly a business pays its suppliers. And shows the amount of times a company pays off its accounts payable in a given period [22].	Turnover
[21]	Average payment period to creditors	365 days / Creditors turnover ratio		Liquidity
[21]	Inventory turnover	Cost of Goods Sold / Average Inventory	It is one of the efficiency ratios that measure how well a business uses its resources. Businesses can improve their decisions on pricing, manufacturing, marketing, and purchasing by using the inventory turnover ratio. Inventory turnover is a financial ratio that demonstrates how frequently a firm turns over its stock in relation to its cost of goods	Turnover

			sold (COGS) over the course of a specific time period [22].	
[21]	Average turnover period for inventories	365 days / Inventory turnover ratio	While divided by the number of days, it is possible to get the average number of days it takes for a firm to sell its inventory [22].	Efficiency
[21]	Gross income divided by sales	Gross Profit / Net Sales or Revenues	The firm's net profit margin, which is calculated by subtracting selling, general, and administrative expenditures, is shown as the gross profit margin [22].	Profitability
[21]	total debt / total assets	Total liabilities / Total Assets	A leverage ratio identified as total-debt-to-total-assets indicates how much debt a corporation has in relation to its assets. Such information may reveal a company's level of financial stability [22].	Leverage
[21]	Net Cash Flow from Operating Activities (NCFOA)	Net Income + Depreciation and Amortization	Net cash flow from operating activites is the amount of money a business earns after deducting taxes and expenses from ongoing, routine business operations, such as producing and selling products or offering clients a service [22].	Liquidity
[21]	operating cash flow / current liabilities	NCFOA / current liabilities	The operating cash flow ratio shows if regular business operations are adequate for a corporation to meet its short-term obligations. A higher ratio indicates that a corporation has produced more cash during a given period than was required to pay off current commitments on an immediate basis [22].	Profitability

[21]	earnings before tax and interest / total asset	EBIT / Total Assets	EBIT divided by total assets financial ratio is regarded as a sign of how successfully a business uses its assets to produce profits [22].	Profitability

In order for a project to be considered successful it does not only need to achieve its goals and objectives but also help the business as a whole. Hence, a project must be linked to an organization which presents healthy and strong financial ratios to succeed. As it is presented in Table 4, a company with favourable liquidity, profitability, and leverage ratios will give the means to finance and carry out profitable projects that will contribute to growth and profitability.

The ability of the business to fulfil its immediate obligations is evaluated by liquidity ratios while leverage ratios determine a company's capacity to fulfil its long-term commitments. A strong ratio from the latter indicate that a company has the funds available to meet its long-term obligations, such as debt repayment or investment in new projects. The third category of ratios is profitability which measure an organization's capacity to make money from its operations. A strong value indicates that a company is making enough money to pay its expenses and make a profit.

Therefore, strong liquidity, profitability, and leverage ratios indicate that a company has the funds available to invest in and carry out profitable projects that will promote growth. Therefore, it is crucial for businesses to keep a focus on their financial ratios and act to raise them if they fall short of standards set by the industry. By doing this, businesses may set themselves up for long-term success and make sure they are prepared to seize new possibilities as they present themselves.

As a result, well-consolidated businesses have favourable financial ratios that show steady financial health. These financial ratios can provide information on matters like the ability to produce a profit and the payment of long- and short-term loans. Making a profit on projects affirms that the organization has the know-how and competence to ensure the success of the project in addition to ensuring that the goals are reached. The submission of ineligible expenses because of not covering all the costs essential for the project's implementation, leading to problems with ineligible expenditure is one of the reasons projects not running successfully.

In conclusion, companies that successfully complete more projects tend to be more prosperous. Therefore, the results of projects are reflected in the company's favourable financial ratios since they not only help the project teams achieve their goals, but also benefit the business.

2.8. Research gap

In the research presented up to this point, and despite our best efforts, no direct study on the prediction of project cancellation was found, in the fields of machine learning and statistical models.

Previous results within this project have pointed to a high correlation between inexperienced enterprises and project cancelation [12]. We believe that there may be a similar relationship between the promoter's financial status and the cancellation of projects applications.

Given that businesses in a poor financial condition may offer ineligible charges and fail to satisfy the proposal initially proposed, the use of ratios used to forecast a company's bankruptcy could be intriguing and helpful in anticipating the project successfulness. Therefore, companies that have a weak financial reputation, characterized by financial ratios, may be more likely to be cancelled afterward.

Consequently, the objective of this work is to create a machine learning model that can forecast whether expenses submitted to funding organizations are eligible or not. It is intended to use the models with better results, namely XGBoost, mentioned in the literature review in order to complement and attempt to improve the model development. It will be verified whether this model exhibits the same behaviour for the improvement in results of accuracy in predicting project successfulness. Instead of employing statistical methods, which are the most common, the employment of more important feature extraction techniques, such as ANOVA F-value, Mutual Information and Feature Importance, is intriguing for the current situation.

The articles that are considered most relevant to assist the intended study are shown in the Table 5:

Table 5. Most relevant articles of literature review.

Article	**Relevance**
[12]	The article that comes closest to the objective that is intended with this dissertation. It shows the following variables as most important: *paid, age, avg_emp* and *y2008*. *Paid* and *age* have been shown to lower the probability of project cancellation. *Age* - number of years of the company - showed that one more year of company consolidation had a positive impact on projects, lowering the probability of project cancellation. *Avg_emp* - average employees - and *y2008* - economic crisis - were shown to have a negative impact on project completion. Implying that a company with a large number of employees and if the project was carried out in a year of economic crisis, the probability of cancellation was higher.
[15]	Features extraction technique different from other statistical methods - *shapley*. This feature extraction demonstrated the ratios that are strongly correlated with a company's bankruptcy. *AUROC* metric was used as evaluation of machine learning models, presented *XGBoost* as a model that increases prediction accuracy. Presenting a 14.81% increase (0.81 to 0.93) when using the XGBoost model compared to the other models applied. Ratios associated with the bankruptcy of a company: *profits before taxes plus interests paid*, and *earnings before income tax and depreciation (EBITDA)*, *trade receivables* and *total assets*. Ratios associated with the non-bankruptcy of a company: *total assets*; *shareholder's funds plus noncurrent liabilities*; *profits before taxes plus interests paid*.
[17]	Demonstration of various results when applying various machine and deep learning models. Bankruptcy prediction is applied for several years (1,2 and 3). The accuracy results of each model decrease as the prediction is made further in advance of the financial crisis. The 3 models with the best results were *RIPPER*, *Local DS* and *C4.5*.
[6]	The inclusion of the cash-flow related ratios present improvements in predicting the bankruptcy of a company. The comparison between

	the results of the two models applied, one of which presents the cash-flow ratio and the other does not, concludes that the financial cash-flow ratio has an influence on predicting the bankruptcy of an organization. As such, the conclusion of this study is in line with Holmen's conclusion " the study of Beaver's univariate model using cash flow/total debt as ratio proved to have a lower error rate compared to the Altman model."

CHAPTER 3

Data preparation and analysis

3.1. CRISP-DM

The methodology used to carry out the dissertation was CRISP-DM (Cross Industry Standard Process for Data Mining), and the various stages of the methodology were followed in order - business understanding, data understanding, data preparation, modelling, evaluation and deployment.

The initial phase, business understanding, was very important for understanding how to approach the project and its objectives and questions. Therefore, a literature review was carried out in order to expand on previous knowledge and studies carried out on the subject.

In the second phase, understanding the data, the various CSV files were collected, and it was initially necessary to explore the multiple files received and understand how they interconnected in order to obtain the feasible dataset. In this phase, more financial data was created according to the information that was already available in the files.

In the third phase, data preparation, once the dataset had been obtained, it was necessary to process and clean the data and select the most important information for the next phase, modelling. More specifically, the most common methods used in this phase were applied, such as treating nulls, identifying outliers and selecting features with the greatest impact on classification. All the methods used in this phase aimed to prepare the dataset for the modelling phase.

For the fourth phase, modelling, various pre-processing experiments, and models were created. The machine learning models trained in this dissertation were: Logistic Regression, Support Vector Classification, Linear Support Vector Classification, K-Nearest Neighbours, Gaussian Naïve Bayes, Perceptron, Decision Tree, Random Forest, Multi-Layer Perceptron Classifier, Extreme Gradient Boost.

In the fifth stage, evaluation, metrics such as accuracy and f1 were used to assess the performance of the experiments created in the previous stage. In addition to this validation, possible optimisations were created to answer the proposed questions.

In this dissertation, the focus was on the data preparation, modelling and evaluation phases in order to obtain a model capable of predicting projects being cancelled or successfully completed.

3.2. Understanding of the problem

The primary aim is to develop a proficient machine learning model that can effectively predict the outcome of projects—whether they will be successfully concluded or cancelled—based on the financial ratios provided during the application process. This is particularly crucial due to the limited availability

of credit for project development. The lending entity's optimal allocation of funds for either initiating new projects or advancing existing ones is of paramount importance.

However, as the aim is to apply artificial intelligence to the data made available and due to the lack of knowledge about analysing applications for project successfulness and financial ratios, a literature review was carried out on these topics in order to come up with a more complete and thorough study.

In the context of this dissertation, projects that have been successfully completed are categorized as closed, with the conclusive criteria being the cessation of investment and project activities. On the other hand, the scope of cancelled projects encompasses various scenarios, such as post-contract cancellations, expiry, and promoter withdrawals. However, for the purpose of this study, only post-contract cancellations are included in the final dataset.

To achieve the objective of identifying a machine learning model capable of differentiating between closed and post-contract cancelled projects based on the financial ratios presented in their respective applications, a comprehensive dataset containing 1356 projects is employed for testing and evaluation.

3.3. Data understanding

Obtaining the final dataset to fulfil the aim of this dissertation was done in several stages, since the final dataset is an aggregation of various files provided by the entity. The three main stages are mentioned below:

- Exclusion of duplicate company registrations for the same year on financial data.
- Combining the two types of existing application files: applications from individuals and organisations.
- Link to closed or cancelled projects for each application.

In the first stage, duplicate information was identified for the same year of company ratios, so it was necessary to eliminate the duplicate records, considering the company record with the most recent date according to when the application was received. The aim of this dissertation is to classify projects at the time of application, so in the second stage a file was added containing information on the applications made by each company and the respective projects. In the third and final stage, to generate the dataset with the necessary information, all closed and cancelled projects were identified according to the files, following the guidelines:

- Closed: all projects that are marked with the project and investment closed. Therefore, records with only one of the closures are not considered closed projects.

- Cancelled: three types of cancellations are presented - post-contract, expiry and withdrawal by the promoter - however, for the purposes of this dissertation, only projects cancelled by post-contract are selected.

In the end, after aggregating the files, the dataset registered 44806 records, but it is not considered the final dataset since it has several records from the same project, but with different years of activity. The ultimate aim of this transformation is for each record to represent the most recent project before the year of the application, in order to simulate the expected cancellation or termination of the project at the time of the application. Therefore, the information and format desired for each project is as follows:

Table 6. Data point structure of project information

Number of the project	Financial ratios inherent in the files	Calculation of financial ratio in relation to 1 year prior to application	Calculation of financial ratio in relation to 2 years prior to application	Calculation of financial ratio in relation to 3 years prior to application

The next step describes the calculation of financial ratios.

3.4. Data preparation

To complete the information in the dataset, it was necessary to calculate the following financial ratios according to the formulae shown in Table 4. These financial ratios will be used later to train the machine learning models. Therefore, the following financial ratios provided by the files were used to carry out the calculations:

- Net Profit for the Period
- Depreciation and Amortization Expenses
- Total Non-Current Liabilities
- Total Current Liabilities
- Total Non-Current Assets
- Total Current Assets
- Sales of Services Rendered
- Current Assets: Accounts Receivable
- Non-Current Assets: Fixed Tangible Assets
- Earnings Before Depreciation and Expenses

After the calculation, the records in the dataset were filtered, reducing the number of records since the aim is to have one record per project according to the most recent year before the application (Table 6). Therefore, each project record continues to have information from previous years of activity, however, they are described through the calculations of the T-1, T-2 and T-3 financial ratios (1, 2 and 3 years before application).

With the dataset in the desired format, it was designed 2 types of datasets according to the presented activity information prior to the year of application:

- Dataset 1 (exclusion of projects): this dataset contains only projects that have information from at least the year prior to the year of application. Therefore, projects with no activity prior to the application are excluded.
- Dataset 2 (no projects excluded): In this dataset no exclusion is made, the projects considered to have no information prior to the application are manipulated so that the values of the financial ratios are 0. There is no loss of data.

In short, the two datasets are used to train the machine learning models to validate the behaviour and results obtained for each one.

3.5. Exploratory data analysis

Dataset 2 (no projects excluded) was expected to have a higher number of records than dataset 1 due to the inclusion of projects with no activity prior to the application, and therefore consists of an increase of 13.28% compared to the first set. Table 7 shows the number of closed and cancelled projects for each dataset:

Table 7. Available datasets and their composition

Dataset excluding projects (1)		
	Count	**Percentage (%)**
Completed	760	63.49
Canceled	437	36.51
Total	1197	100.00
Dataset without excluding projects (2)		
	Count	**Percentage (%)**
Completed	804	59.29
Canceled	552	40.71
Total	1356	100.00

Table 8. Composition of dataset including all projects

Dataset without excluding projects (2)	
Type of projects	Number of projects with no activity prior to the year of application included in dataset 2
Canceled	115
Completed	44

According to the Table 8, 115 cancelled and 44 completed projects were added to the dataset 2. As can be seen from Table 7, the disparity between project types decreases in the second dataset, leading to a more balanced dataset. Therefore, in dataset 2 the completed projects represent 59.29% of the final set compared to dataset 1 which represents 63.49%.

3.6. Outliers

To identify whether there are many disparate values, outliers were identified using two statistical methods – z-score and interquartile range – within dataset 1. These methods are important for identifying data points that deviate considerably from the rest of the data and for determining how far a data point is from the centre of the data distribution.

- Z-score: The distance between a single data point and the mean (average) of all the data points in a group. It aids in determining whether a certain data point is indeed distinct from the rest. A higher z-score indicates that the data point is more out of the ordinary.
- Interquartile range (IQR): By concentrating on the middle 50% of data values, the interquartile range (IQR) is a reliable statistical measure used to characterise the spread or variability within a dataset. The IQR is a measurement of the dispersion of data within the centre region of the distribution and is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). It efficiently summarises the dataset's variability and is highly resistant to the influence of extreme outliers.

Table 9. Number of outliers identified by interquartile range

Interquartile Range	
Number of outliers per datapoint	Number of datapoints to disregard
5	1073
10	824
15	621

For the Z-score method there were 267 features which exceeded a threshold of 2 (Table 35 in annexes) while the method IQR identified the maximum of 1073 datapoints as outliers if identified 5 outliers per datapoint and the minimum of 621 datapoints as outliers if using 15 outliers per datapoint (Table 9).

Despite the existence of outliers in the dataset as seen in Table 9 and Table 35, the removal of outlier records was not applied since it was intended for the trained models to be able to predict completed projects from cancelled ones by presenting values that are far from the expected ones and the dataset is not large enough to be excluding more records.

3.7. Feature extraction

As the aim of the dissertation is to focus on a classification problem, methods have been used to extract the features that have the greatest impact on the classification of a project that will end or be cancelled. By identifying the best features, it was possible to see which types of information have the greatest or least impact on the results of the models. The following list shows the feature selection methods used:

- ANOVA F-value: This statistical technique assesses the variation in means across different classes or groups within a target variable. When coupled with a low p-value, a high F-value indicates that the associated features play a significant role in distinguishing between these classes.
- Mutual Information: Mutual information quantifies the degree of interdependence between two variables. When utilized for feature selection, it evaluates the amount of information one feature can provide about another.
- Feature Importance with Decision Trees, Random Forests, and XGBOOST: This approach gauges the significance of each feature in reducing impurity (such as Gini impurity or entropy) at decision nodes within these tree-based models.

Before executing the feature selection methods, values which presented nulls and infinites values were replaced for 0. After that treatment of nulls, features, whose values were constant were extracted. Based on this, 76 columns were removed from the dataset (Table 38, annex pp.97).

After the treatment of features, which presented null and infinite values, and removing values which were constant, the next step consisted of extracting the features that had a stronger impact on project cancellation, using the methods of feature selection mentioned above. For this reason, the application of the methods had access to all the features available from csv files, except for the features already removed from Table 38, in annex pp.97. For each method the top 30 and 50 features of the dataset were selected, representing 11% and 18% of the available features,

respectively. With these, the purpose was to test that with a reasonable number of features the experiments could achieve more interesting results. Another aim was to verify how the model would behave when utilizing either the entire dataset's features or exclusively the financial ratios manually computed in accordance with the literature review based on Table 4.

The following Table 10 shows a resume of all the list of features selected for the next experiments, the number of features used to train the models according and the method used for feature selection:

Table 10. Features and extraction method

Name of the list	**Extraction method**	**Number of features selected**	**Features**
feat1	ANOVA F-value	30	Feature set – feat1 (Table 35 – annex, pp.83)
feat2	Mutual information	30	Feature set – feat2 (Table 35 – annex, pp.83)
feat3	ANOVA F-value	50	Feature set – feat3 (Table 35 – annex, pp.83)
feat4	Mutual information	50	Feature set – feat4 (Table 35 – annex, pp.84)
feat5	Decision Tree	50	Feature set – feat5 (Table 35 – annex, pp.84)
feat6	Random Forest	50	Feature set – feat6 (Table 35 – annex, pp.85)
feat7	XGBoost	50	Feature set – feat7 (Table 35 – annex, pp.85)
allFeat	Manual	267	Feature set – allFeat (Table 35 – annex, pp.86)
manualFeat	Manual	22	Feature set – manualFeat (Table 35 – annex, pp.88)

For each list built with algorithmic extraction method – feat1, feat2, feat3, feat4, feat5, feat6 and fea7 – there is at least one financial ratio within. The following Table 11 describes the number of financial ratios for each feature list and the number of ratios categories:

Table 11. Financial ratios in each feature list

Name of the list	Number of financial ratios	Categories
feat1	4	Profitability - 2 Efficiency - 1 Overall financial ratio - 1
feat2	1	Profitability - 1
feat3	4	Profitability - 2 Efficiency - 1 Overall financial ratio - 1
feat4	4	Profitability - 1 Leverage - 1 Efficiency - 1 Overall financial ratio - 1
feat5	9	Profitability - 5 Leverage - 1 Coverage - 1 Turnover - 1 Overall financial ratio - 1
feat6	13	Profitability - 7 Leverage - 1 Efficiency - 2 Coverage - 1 Liquidity - 1 Overall financial ratio - 1
feat7	8	Profitability - 5 Efficiency - 2 Coverage - 1

CHAPTER 4

Discussion of results

4.1. Experiences

Four different test scenarios were devised, using the 2 data sets referenced in Table 7. The data set 2 referenced in Table 7 does not involve any filtering for projects that do not have information on commercial activity in the previous year, while data set 1 only contains projects that have commercial activity in the previous year. In short, data set 1 is summarised as *commercially active projects* and data set 2 as *all projects*. Each scenario has different characteristics regarding the data under consideration. Scenarios 1, 2.a and 2.b are based on as *commercially active projects* data set (data set 1 in Table 7). However, scenarios 2.a and 2.b implement filtering at the level of cancelled projects, as shown in Table 12. Scenario 3, on the other hand, is based on *all projects data set* (data set 2 in Table 7).

For each scenario, it was select seventy percent of the data to train the machine learning model, while the rest of the thirty percent of data was used for test reasons to verify the performance of the trained model.

Table 12 below provides a concise summary of the four test scenarios, including the data source used (Table 7), a brief description of the considered data and the total records count for each respective scenario:

Table 12. Description of test scenarios

Scenario	Used dataset	Considered data	Number of records
1	Commercially active projects	Projects completed and cancelled post-contract, filtering projects that did not present activity data from previous years in relation to the year of application.	Training records: 838 Testing records: 359 Total: 1197
2.a	Commercially active projects	Completed projects and projects cancelled by post-contract, where the reason was "Promoter's withdrawal".	Training records: 684 Testing records: 293 Total records: 977
2.b	Commercially active projects	Completed projects and projects cancelled by post-contract,	Training records: 686 Testing records: 284 Total records: 980

		where the reason is not "Promoter's withdrawal"	
3	All projects	All valid projects to be classified as completed and cancelled (post-contract)	Trained records: 949 Tested records: 407 Total records: 1356

For each proposed scenario, i.e., for each type of data set, a set of tests was applied based on the combination of an ML model and a list of features (Table 10). The most common machine learning models used for classification problems were trained, and then a comparison was made to obtain the model with best results. Therefore, the following 10 models were used:

- Logistic regression
- Support vector classification
- Linear support vector classification
- K-nearest neighbors
- Gaussian naïve bayes
- Single layer perceptron, also known as, perceptron
- Decision tree
- Random forest
- Multi-layer perceptron classifier
- Extreme gradient boost

All the parameters used to train the models above are mentioned in Table 37 in the annex.

From this combination ML model and list of features, 5 pre-processing experiments are carried out, as mentioned in Table 13. The purpose of this pre-processing experiments was to conclude the impact of using standard scaler, principal component analysis (PCA) and balancers as synthetic minority oversampling technique (SMOTE) and random under sampling. So, each test or experiment mentioned in the dissertation is made up of a model, a list of features and a pre-processing experiment.

To verify which lists of features, models and pre-processing experiments had the best results, several tests were carried out. Each test consisted of using one model listed above, one list of features (mentioned in Table 10) and a pre-processing experiment. Other objectives to achieve were to verify if the number of the features, the type of features (using all features available from the dataset or selecting only the financial ratios calculated manually as shown in Table 10.) had impact in the improvement of the metrics.

Therefore, for each scenario mentioned in Table 12, 150 experiments are carried out, totalling 600 experiments executed over the course of the dissertation. The Table 13 below describes the 5 types of pre-processing that make up the experiments.

Table 13. Description of pre-processing experiments

Pre-processing experience	Description
Baseline	No processing of any kind is carried out on the data set. The models are trained according to the values in the files received.
Pre-processing with standard scaler	All the values in the data set are to scale, with no scattered values.
Pre-processing with standard scaler plus PCA	Scaled dataset records and use of PCA for dimensionality reduction.
Pre-processing with standard scaler plus SMOTE	Scale dataset values and use SMOTE to generate synthetic records of the class with the fewest records – cancelled projects.
Pre-processing with standard scaler plus SMOTE and random under sampling	Scale dataset values and use SMOTE to generate synthetic records from the class with the fewest records and random under sampling to remove some records from the largest class. With the aim of balancing the training dataset as much as possible.

After conducting 600 experiments, individual reports were created for each scenario, detailing the metrics acquired in each experiment. Consequently, the following metrics were employed within the scope of the analysis to evaluate and categorize the experiment demonstrating the highest proficiency in distinguishing closed projects from cancelled ones:

1. f1: Balances precision and recall for a trade-off between false positives and false negatives.
2. ROC area under curve: Quantifies a model's ability to distinguish classes using the ROC curve area.
3. Precision: emphasizes the accuracy of positive predictions. It measures the proportion of true positive predictions among all instances that the model predicted as positive.
4. Recall: emphasizes the model's ability to capture all positive instances. It measures the proportion of true positive predictions among all actual positive instances.
5. Accuracy: Measures correct predictions relative to the total instances.

It is important to emphasise that the f1 metric was a more important metric for finding the experiences, models and features that had the greatest impact on the classification problem being the preferred metric for evaluating the performance of classification models. Its strength lies in its ability to accurately assess model performance on both balanced and imbalanced datasets.

In order to make it easier to analyse the results, experiments in which the model exhibited overfitting and those yielding accuracy results lower than the scenario-specific dummy values were excluded from the report. The dummy results from each scenario were calculated through the strategy "most frequent" which predicts the class label that appear most frequently given each scenario dataset.

Within the 4 scenarios, in the first 3 there were experiments that overfitted with the models Random Forest, Decision Tree, Multi-Layer Perceptron Classifier and Extreme Gradient Boosting. However, the Random Forest and Decision Tree models were the only ones that were completely excluded in the reporting of scenarios 1 and 2 (version a and b) because all the experiments carried out resulted in overfit in the training models.

The "Experiences removed" column in Table 14 refers to the number of experiments removed from each scenario for the reasons mentioned above (overfitted and underperformed models). The following Table 14 describes the accuracy of the dummy for each scenario, the number of models that overfitted, the number of models that obtained results lower than the accuracy of the dummy classifier, the total number of experiments removed and the machine learning (ML) models that were removed completely from the analysis, without any experiments to analyse:

Table 14. Filter of overfit and results below dummy scores

Scenario	Dummy Accuracy	Overfit	Less than dummy result	Experiences removed	Models excluded (completely)
1	0.63	133	118	220	Random Forest, Decision Tree
2.a	0.78	131	290	299	Random Forest, Decision Tree
2.b	0.59	129	213	290	Random Forest, Decision Tree
3	0.78	0	32	32	ND

With the second scenario, it was validated that there were more experiences that obtained poorer results than the dummy result. Therefore, the exclusion of more projects resulted in the degradation of results in certain models and experiments, increasing the number of discarded experiments. In the fourth scenario (scenario 3), the inclusion of a greater number of projects resulted in benefits for the training of the models as they were not overfitted. Therefore, the inclusion of projects that were initially excluded from scenario 1 (filtered projects that did not have activity information before the year of application) helped to avoid overfitting the Random Forest and Decision Tree models, which are predisposed to do so. □

4.2. Count of pre-processing experiences, list of features and projects

After filtering out experiments with accuracies lower than the results of the dummies or which resulted in an overfitted model, the experiments were counted according to each pre-processing method and the list of features used in each scenario 1, 2.a, 2.b and 3. The figures listed below (Figure 2, Figure 3, Figure 4 and Figure 5) reflect the counts for the two themes.

In the 4 scenarios presented, there are fewer experiments using only the *manualFeat* list (calculated financial ratios), with a noticeable difference in scenario 2.a where there were only 2 experiments selected. The frequency of experiments with the features from list 7 in scenario 2.a is also very low.

Regarding the pre-processing experiments, experiments 2 (standard scaler) and 3 (standard scaler and principal component analysis - PCA) in scenarios 1 and 2 are the experiments with the greatest emphasis and it was found that they were the ones that obtained the fewest results biased or inferior to the dummy models, reaching a greater count of experiences as seen in Figure 2, Figure 3 and Figure 4. However, the frequency of experiments 1 (baseline) is lower compared to the others, suggesting that applying pre-processing such as standard scaler and principal component analysis (PCA) could help improve the results. It was also found that in scenario 2, overall, there were few experiments with synthetic minority oversampling technique (SMOTE) and random under sampling, suggesting that most of the results from these experiments were below the criteria of the dummy model or overfitted.

Figure 6 shows the count of completed and cancelled projects according to the number of years of activity available for consultation prior to application. There is a higher concentration of cancelled projects when there is no information on any previous activity. A greater number of completed projects have between 1 and 4 years of information prior to application. In cases where there is information for 5 or more years, many projects are cancelled.

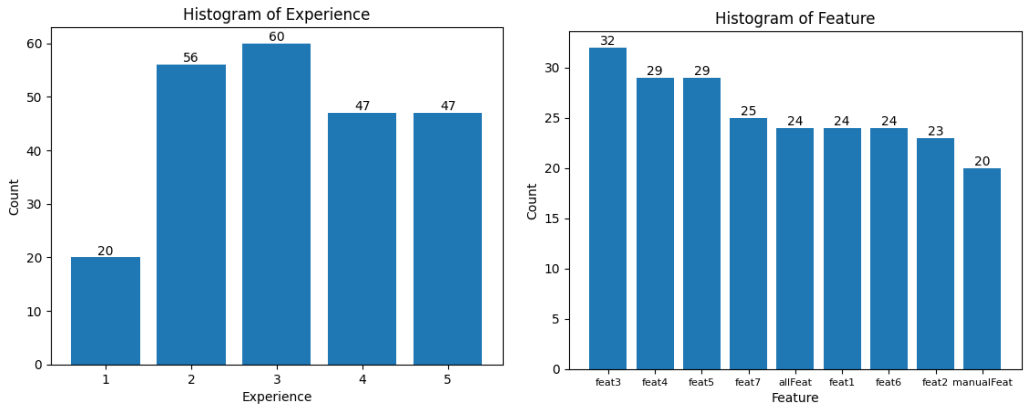

Figure 2. Count of experiences of pre-processing and list of features from scenario 1

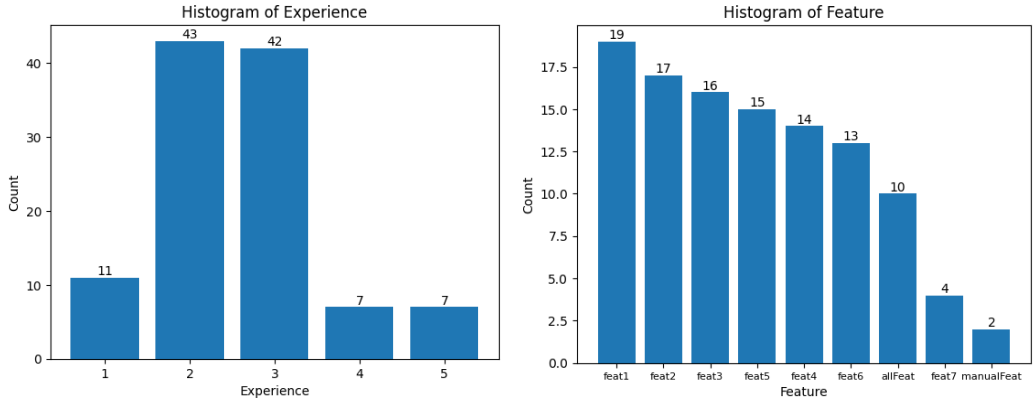

Figure 3. Count of experiences of pre-processing and list of features from scenario 2.a

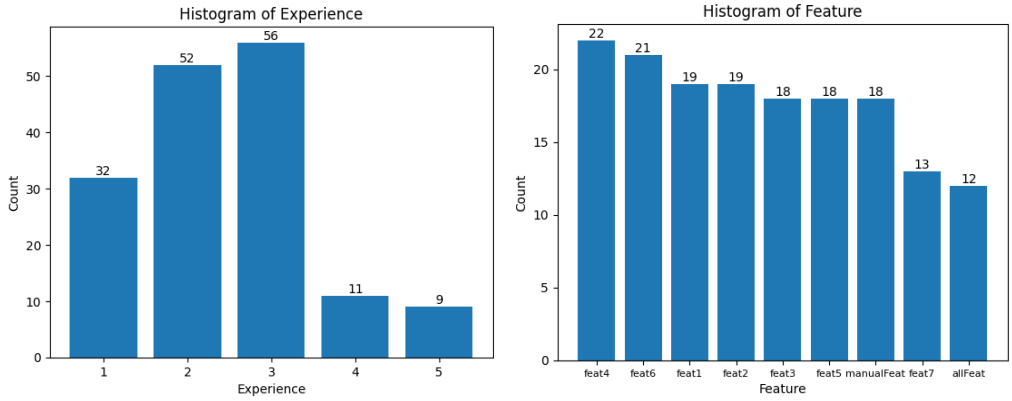

Figure 4. Count of experiences of pre-processing and list of features from scenario 2.b

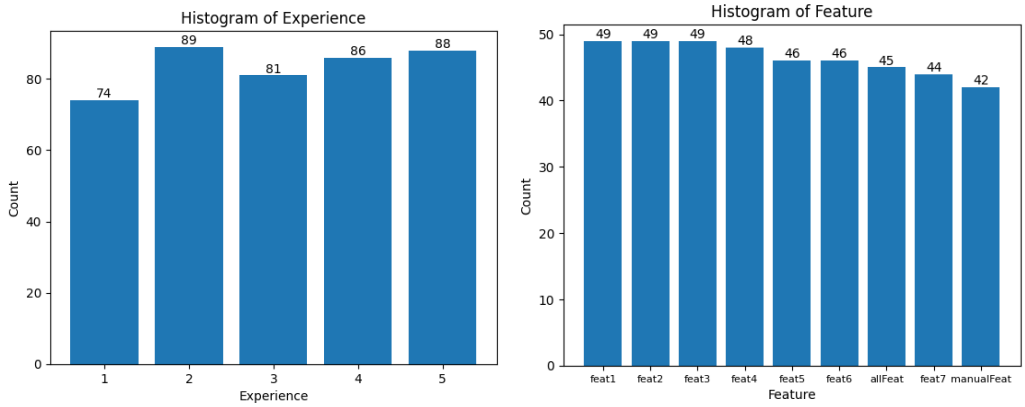

Figure 5.Count of experiences of pre-processing and list of features from scenario 3

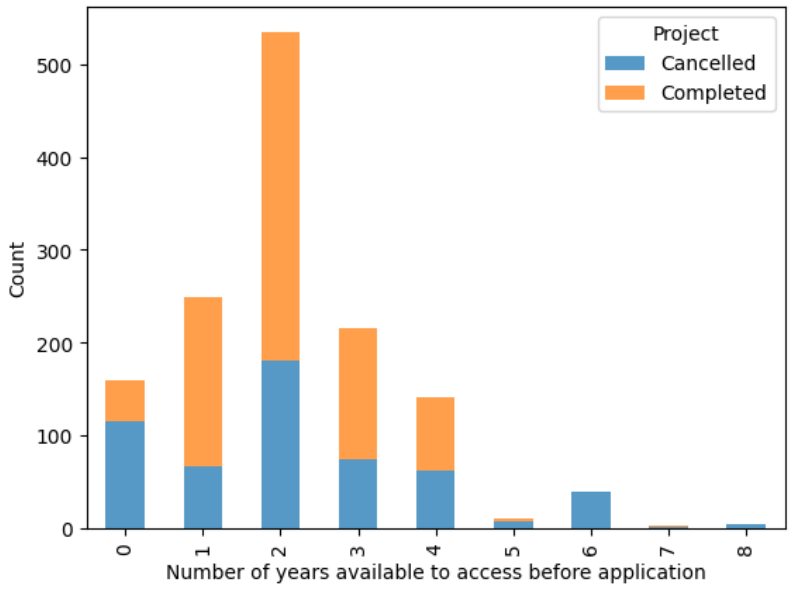

Figure 6. Count of cancelled and completed projects according to number of years available

4.3. Best f1 results by model

Of the various experiments carried out, the trained models were grouped together, and the best f1 results obtained for each test scenario. The Table 15 below shows the results for each model in each scenario:

Table 15. Best f1 result from each model by scenario

	Scenario 1	**Scenario 2.a**	**Scenario 2.b**	**Scenario 3**
Gaussian NaiveBayes	0.52	0.34	0.34	0.63
K-NearestNeighbors	0.60	0.43	0.48	0.65
LinearSupportVectorClassification	0.62	0.40	0.34	0.65
LogisticRegression	0.61	0.46	0.44	0.67
MLPClassifier	0.61	0.44	0.62	0.66
Perceptron	0.59	0.43	0.48	0.67
SupportVectorClassification	0.66	0.31	0.58	0.65
XGBoost	0.58	0.42	0.43	0.65
RandomForest	ND	ND	ND	0.65
DecisionTree	ND	ND	ND	0.63

For each of the scenarios 1, 2 (scenario a and b), and 3, the average f1 score is as follows: 0.60, 0.40, 0.46, and 0.65, respectively.

- Analysis Scenario 1: The average f1 score from scenario 1 – 0.60 – compared to scenario 2 (a and b) – 0.40 and 0.46 – is far superior, which means that the results of the models trained in scenario 1 are greater than scenario 2, with the exception of the Multi-Layer Perceptron (MLP) in scenario 2.b which had the same result in comparison to the same model in scenario 1.
- Analysis Scenario 2.a: Although no model obtained an f1 result equal to or greater than 0.5, the one that came closest was the Logistic Regression model with a value of 0.46. The models in this scenario had the worst f1 results, reflected in the average f1 result of 0.40.
- Analysis Scenario 2.b: The results of the models in this scenario are similar to scenario 2.a, with f1 results below 0.5, with the exception of the Multi-Layer Perceptron (MLP) and Support Vector Classification (SVC) models, which obtained interesting values - 0.62 and 0.58 respectively.
- Analysis Scenario 3: The majority of the models had superior results in comparison to the rest of the scenarios, in exception for the model Support Vector Classification (SVC) of scenario 1 which had a superior result by a small margin. In comparison to the scenario 2, the results of scenario 3 are far greater.

In short, the f1 results obtained in scenario 2 were lower than any other scenario tested, but this behaviour may be due to a greater imbalance between the types of projects. It is important to emphasise that only the scenario 3 had all the models trained, and the models excluded in other scenarios (1 , 2.a and 2.b) – Random Forest (RF) and Decision Tree (DT) – obtained similar f1 results in comparison to the other models in the respective scenario.

4.4. Models and experiences

In this section the best f1 result was extracted. This result was achieved by the combination of model and pre-processing experiment. Analysing the f1 results for the combination of model and experiment, it was observed that the f1 results for scenarios 1 and 3 are higher than the scenarios 2.a and 2.b, which presents an average of 0.39 and 0.38, respectively. Scenarios 1 and 3 have f1 values of 0.54 and 0.62 respectively. By observing the results of the combination of model and experiment, the following important characteristics and behaviours can be drawn:

- Each model presented in each scenario have the experiences 2 and 3, with the exception of the XGBoost model in scenarios 1 and 2, where they only had the experiment 3 (Table 27, Table 29, Table 31 and Table 33 – annex).
- The majority of models which carried generation of synthetic data tests improved the f1 results. However, the only model which did not have that behaviour was Gaussian Naïve Bayes (GNB) from the scenarios 2.b and 3.

- Most of the experiments from each model which did not have any type of pre-process – baseline - had the worst results in the respective model.

Below is a list of observations on the results of each scenario and their best combination of model and pre-processing experience according to the Table 16:

- The models in scenario 1 that obtained the best f1 results were Support Vector Classification (SVC) and Linear Support Vector Classification (LSVC). In this scenario there were 5 models and only 1 of them failed, Gaussian Naïve Bayes (GNB), to obtain experiments with results higher than 0.54 – the average of the accuracy results for scenario 1.
- The average f1 results for each version (a and b) of scenario 2 are 0.39 and 0.38, respectively. However, the highest result for each version is far apart, with version *a* obtaining 0.46 from Logistic Regression (LR) compared to 0.62 from Multi-Layer Perceptron (MLP) from version b. For scenario 2.a, the models Support Vector Classification (SVC) and Gaussian Naïve Bayes (GNB) do not have experiments which f1 results are equal or greater than the average result of the scenario. For the scenario 2.b, the models which do not have any test that obtained f1 result greater or equal to the average are Gaussian Naïve Bayes (GNB) and Linear Support Vector Classification (LSVC).
- In the scenario 3, all models from the scenario had at least 1 experiment which obtained f1 result greater than 0.62 – the average of the f1 results for scenario 3. The models Logistic Regression (LR) and Single Layer Perceptron (SLP) with the experiences 4 and 5 obtained the best result from all the scenarios, 0.67.

Table 16. Best f1 results with the combination model and pre-processing experience from each scenario

Model	**Pre-processing experience**	**F1**
	Scenario 1	
Support Vector Classification	Standard scaler + SMOTE	0.66
Support Vector Classification	Standard scaler + SMOTE + random under sampling	0.66
Linear Support Vector Classification	Standard scaler + SMOTE + random under sampling	0.61
	Scenario 2.a	
Logistic Regression	Standard scaler	0.46
Logistic Regression	Standard scaler + PCA	0.46
Multi Layer Perceptron (MLP)	Standard scaler	0.44

Scenario 2.b		
Multi Layer Perceptron (MLP)	Standard scaler + SMOTE	0.62
Multi Layer Perceptron (MLP)	Standard scaler + SMOTE + random under sampling	0.62
Support Vector Classification	Standard scaler + SMOTE	0.58
Support Vector Classification	Standard scaler + SMOTE + random under sampling	0.58
Scenario 3		
Logistic Regression (LR)	Standard scaler + SMOTE	0.67
Logistic Regression (LR)	Standard scaler + SMOTE + random under sampling	0.67
Single Layer Perceptron (SLP)	Standard scaler + SMOTE	0.67
Single Layer Perceptron (SLP)	Standard scaler + SMOTE + random under sampling	0.67
Multi Layer Perceptron (MLP)	Standard scaler + SMOTE	0.66
Multi Layer Perceptron (MLP)	Standard scaler + SMOTE + random under sampling	0.66

Since the dataset was not balanced, Table 17 shows the models that benefited from the synthetic data generation experiment when training the data set and obtained the best f1 results compared to the other experiments in the same model. Scenario 3 shows the positive impact of this synthetic data generation experiment due to the large number of models – Gaussian Naive Bayes, K-Nearest Neighbours, Linear Support Vector Classification, Logistic Regression, Multi-Layer Perceptron, Single Layer Perceptron, Random Forest, Support Vector Classification – that obtained better results by balancing the data set. The pre-processing experience 4 used standard scaler plus smote, while pre-processing experience 5 added random under sampling to the latter experience.

Table 17. Models benefited from experiences with SMOTE in each scenario

Model	**Pre-processing experience**	**F1**
Scenario 1		
K-Nearest Neighbours	4, 5	0.60
Linear Support Vector Classification	5	0.62
Logistic Regression	4, 5	0.61
Multi Layer Perceptron	5	0.59

Support Vector Classification	4, 5	0.66
Scenario 2.a		
Multi Layer Perceptron	4, 5	0.62
Support Vector Classification	4, 5	0.58
Scenario 2.b		
Multi Layer Perceptron	4, 5	0.62
Support Vector Classification	4, 5	0.58
Scenario 3		
Gaussian Naive Bayes	4, 5	0.63
K-Nearest Neighbours	4	0.65
Linear Support Vector Classification	4, 5	0.65
LogisticRegression	4, 5	0.67
Multi Layer Perceptron	4, 5	0.66
Single Layer Perceptron	4, 5	0.67
RandomForest	4, 5	0.65
SupportVectorClassification	4, 5	0.65
Gaussian Naive Bayes	4, 5	0.63

4.5. Metrics of each feature list and model

Two analysis were carried out in relation to the metrics, one at the level of the features used and the other at the level of the models used. The median was applied to each section - features and model - so Tables 18 and 19 provide the median of the metrics - accuracy, f1, precision, recall, receiver operating characteristics area under curve – obtained in the experiences for each list of features and model tested for each scenario. At feature level, the following observations based on Table 18 were made for each scenario:

- In scenario 1, the *allFeat* featureList showed the best results for the metrics: Accuracy, f1, Precision, Recall, Receiver Operating Characteristics Area Under Curve (ROC_AUC). The *feat2* feature list showed the best precision in scenario 1, a large value with a notable disparity in relation to the other feature lists presented in the same scenario. However, the f1, Recall and Receiver Operating Characteristics Area Under Curve (ROC_AUC) metrics are among the worst within the scenario, emphasising that the values presented are relatively lower than the others.
- For the scenario 2.a, the feature lists with the best f1 results were *manualFeat* and *feat5* with values of 0.43 and 0.40, respectively. However, it should be noted that the list *manualFeat*

only had 2 experiments for scenario 2.a. Therefore, as the sample is so small, it is not possible to conclude with certainty that the list of calculated financial ratios is better than the other lists. The *feat2* list has the best precision (0.82), in contrast to the *manualFeat* list which has the worst result (0.59), but the recall was the highest from other features lists, making the *manualFeat* with higher f1 score. The list which had the worst performance according to the criteria of f1 was *feat7*, presenting a low value of f1 of 0.26. Regarding scenario 2.b, the best f1 results are presented by the *feat5* list of features and the worst in the scenario are found in the *feat1*, *feat2* and *allFeat* with value of 0.22. It should be noted that in both versions of scenario 2 the presented f1 results are below from expected, being lower than 0.5.

- In the scenario 3, the best performing features are in the *feat4* list, with 4 metrics with the best results – Accuracy, f1, Precision, Recall, ROC_AUC. And with regard to the worst results, the *manualFeat* were the ones that showed the worst results – Accuracy, f1, Precision, Recall, ROC_AUC – however, it is important to note that these results are not very far from the rest of the test values. All the features applied in this scenario have f1 results higher than 0.5.

Table 18. Metrics of features list from each scenario

Features_List	Accuracy	F1	Precision	Recall	ROC_AUC
Scenario 1					
allFeat	0.68	**0.56**	0.62	**0.52**	**0.65**
feat4	0.68	0.51	0.64	0.42	0.63
feat5	0.67	0.51	0.63	0.48	0.62
feat7	0.66	0.49	0.6	0.4	0.62
feat1	0.66	0.48	0.64	0.39	0.62
manualFeat	0.65	0.47	0.63	0.37	0.61
feat6	0.68	0.46	0.63	0.35	0.62
feat3	0.67	0.44	0.68	0.34	0.61
feat2	0.66	0.25	**0.82**	0.16	0.56
Maximum value	0.68	0.56	0.82	0.52	0.65
Average value	0.67	0.46	0.65	0.38	0.62
Minimum value	0.65	0.25	0.60	0.16	0.56
Scenario 2.a					
manualFeat	0.79	0.43	0.59	**0.34**	**0.63**
feat5	0.80	0.40	0.75	0.26	**0.62**
allFeat	0.80	0.37	0.80	0.24	0.61

feat2	0.80	0.32	**0.82**	0.20	0.59
feat1	0.80	0.32	0.78	0.20	0.59
feat4	0.80	0.32	0.77	0.21	0.59
feat6	0.79	0.32	0.68	0.21	0.59
feat3	0.80	0.32	0.78	0.20	0.59
feat7	0.78	0.26	0.69	0.16	0.57
Maximum value	0.80	0.43	0.82	0.34	0.63
Average value	0.79	0.34	0.74	0.23	0.60
Minimum value	0.78	0.26	0.59	0.16	0.57
Scenario 2.b					
feat5	0.80	0.35	0.61	0.25	**0.60**
feat3	0.81	0.32	0.82	0.20	**0.59**
feat6	0.80	0.32	0.60	0.20	**0.59**
feat7	0.78	0.28	0.57	0.20	0.57
feat4	0.80	0.26	0.68	0.16	0.57
feat1	0.80	0.22	**0.89**	0.12	0.56
feat2	0.80	0.22	**0.89**	0.12	0.56
allFeat	0.80	0.22	0.78	0.12	0.56
manualFeat	0.78	0.09	0.53	0.05	0.52
Maximum value	0.81	0.35	0.89	0.25	0.60
Average value	0.80	0.25	0.71	0.16	0.57
Minimum value	0.78	0.09	0.53	0.05	0.52
Scenario 3					
feat4	0.68	0.60	0.67	0.56	0.67
feat5	0.67	0.60	0.65	0.59	0.66
allFeat	0.67	0.60	0.64	0.58	0.66
feat3	0.68	0.60	0.66	0.57	0.65
feat6	0.66	0.60	0.66	0.54	0.65
feat7	0.66	0.60	0.63	0.58	0.65
feat1	0.67	0.59	0.66	0.54	0.65
feat2	0.66	0.59	0.65	0.53	0.65
manualFeat	0.65	0.58	0.65	0.53	0.63
Maximum value	0.68	0.60	0.67	0.59	0.67

Average value	0.67	0.60	0.65	0.56	0.65
Minimum value	0.65	0.58	0.63	0.53	0.63

Below are the observations made at model level for each scenario based on Table 19

1. The GNN, MLP, and SLP models stood out as the most noteworthy performers, primarily due to their f1 scores exceeding 0.50. Among these three models, the GNN model demonstrated the most promising outcomes within this context, achieving an accuracy of 0.68, an f1 score of 0.56, a precision score of 0.62, a recall score of 0.52, and a roc_auc score of 0.65. Meanwhile, the remaining models in this scenario exhibited f1 scores of 0.44 or higher, rendering them viable candidates. However, the LSVC model's f1 score of 0.25 indicates poor performance and should be excluded from consideration.
2. In scenario 2, both versions, a and b, exhibit a noteworthy characteristic in their accuracy results. These models consistently achieved high accuracy scores, maintaining an average accuracy of 0.80 across both versions. Nevertheless, none of the models in this scenario achieved an f1 score equal to or greater than 0.5. The model that came closest to this threshold was the MLP in scenario 2.b. Consequently, within scenario 2, it is the MLP model that stands out as the top performer, especially in terms of f1 score, with a value of 0.43. Despite demonstrating high precision values, the models in scenario 2 faced a significant challenge with low recall values, which had a substantial adverse impact on the f1 score. Scenario 2.a achieved an average recall of 0.21, while scenario 2.b reached 0.20.
3. In scenario 3, a majority of the models listed in Table 19—namely MLP, XGBoost, RF, SLP, DT, SVC, LSVC, KNN, and LR—exhibited f1 scores ranging from 0.58 (for LR, KNN, and LSVC) to a peak of 0.62 (achieved by MLP). However, it's worth noting that the GNB model displayed an f1 score of 0.19, primarily due to its exceptionally low recall value of 0.11. In terms of f1 performance, the standout model within scenario 3 was unequivocally MLP.

For both the analysis of the features and the models, the following behaviour can be seen: the results presented in scenario 3 are the “best” since there is no trade-off between the various metrics and the scores of f1 presented are the highest between the scenarios.

All the metric values in the scenario 3 seem to be in agreement and within a similar range of values. Whereas in scenario 2 there are better accuracies but at the cost of a very low recall.

Table 19. Metrics of models from each scenario

Model	Accuracy	F1	Precision	Recall	ROC_AUC
Scenario 1					
Gaussian NaiveBayes	0.68	0.56	0.62	0.52	0.65
MLP Classifier	0.68	0.51	0.64	0.42	0.63
Perceptron	0.67	0.51	0.63	0.48	0.62
XGBoost	0.66	0.49	0.60	0.40	0.62
K-Nearest Neighbors	0.66	0.48	0.64	0.39	0.62
Support Vector Classification	0.68	0.46	0.63	0.35	0.62
Logistic Regression	0.67	0.44	0.68	0.34	0.61
Linear Support Vector Classification	0.66	0.25	0.82	0.16	0.56
Maximum value	0.68	0.56	0.82	0.52	0.65
Average value	0.67	0.46	0.66	0.38	0.62
Minimum value	0.66	0.25	0.60	0.16	0.56
Scenario 2.a					
Perceptron	0.80	0.40	0.75	0.26	0.62
Gaussian NaiveBayes	0.80	0.37	0.80	0.24	0.61
Linear Support Vector Classification	0.80	0.32	0.82	0.20	0.59
K-Nearest Neighbors	0.80	0.32	0.78	0.20	0.59
Logistic Regression	0.80	0.32	0.78	0.20	0.59
MLP Classifier	0.80	0.32	0.77	0.21	0.59
Support Vector Classification	0.79	0.32	0.68	0.21	0.59
XGBoost	0.78	0.26	0.69	0.16	0.57
Maximum value	0.80	0.40	0.82	0.26	0.62
Average value	0.80	0.33	0.76	0.21	0.60
Minimum value	0.78	0.26	0.68	0.16	0.57
Scenario 2.b					
MLP Classifier	0.8	0.43	0.56	0.35	0.64
XGBoost	0.79	0.39	0.55	0.30	0.61
K-Nearest Neighbors	0.8	0.38	0.62	0.25	0.61
Perceptron	0.79	0.30	0.52	0.20	0.58
Logistic Regression	0.80	0.27	0.69	0.17	0.57
Linear Support Vector Classification	0.80	0.24	0.67	0.15	0.56
Gaussian NaiveBayes	0.81	0.22	1	0.12	0.56
Support Vector Classification	0.79	0.11	1	0.06	0.52

Maximum value	0.79	0.43	0.52	0.06	0.52
Average value	0.80	0.29	0.70	0.20	0.58
Minimum value	0.81	0.11	1.00	0.35	0.64
Scenario 3					
MLPClassifier	0.69	0.62	0.65	0.60	0.68
XGBoost	0.68	0.61	0.66	0.57	0.66
RandomForest	0.70	0.60	0.70	0.52	0.67
Perceptron	0.63	0.60	0.60	0.64	0.63
DecisionTree	0.63	0.59	0.58	0.61	0.63
SupportVectorClassification	0.67	0.59	0.67	0.57	0.65
LinearSupportVectorClassification	0.67	0.58	0.63	0.59	0.66
K-NearestNeighbors	0.66	0.58	0.64	0.52	0.65
LogisticRegression	0.67	0.58	0.65	0.50	0.66
GaussianNaiveBayes	0.60	0.19	0.95	0.11	0.55
Maximum value	0.69	0.62	0.70	0.64	0.68
Average value	0.67	0.60	0.65	0.56	0.66
Minimum value	0.60	0.19	0.58	0.11	0.55

4.6. Best experience from each scenario

Considering the final report with all the conditions and experiences taking into account and sorting to obtain the best combination of model, experience and features list, the metrics are sorted in the following order: f1, roc_auc, precision, recall and accuracy.

After applying a second filter on results - eliminating tests in which the f1 values are lower than the scenario average – there are about around half of the tests in which the f1 values are higher than the reporting average, this behaviour is common between the 4 scenarios.

The following Table 20 shows the total number of tests in each scenario and the number of tests with an accuracy higher than the scenario average:

Table 20. Number of tests greater or equal to the average f1 of scenario

Scenarios	**Total number of tests after first filtering**	**Number of tests after second filter**	**Average f1**
1	230	133 (57.83%)	0.46
2.a	110	39 (35.45%)	0.34
2.b	321	297 (71.05%)	0.34
3	418	184 (57.32%)	0.56

Therefore, supported on the experiences in which f1 is equal to or higher than the average for each scenario, the following results were obtained:

- In the tests carried out in the first scenario, there were 133 tests that obtained an f1 higher than the average f1 value for the scenario – 0.46. The model and experiments that obtained the best f1 result (0.66) was SVC with the experiments 4 and 5 – Standard Scaler with SMOTE and Standard Scaler with SMOTE plus random under sampling – with the *feat5* feature list. With these experiences it achieved the following values 0.72, 0.63, 0.70, 0.72 for accuracy, precision, recall and roc_auc, respectively (Table 21). The worst experience was presented by the GNN model with the experiment 3 (standard scaler plus principal component analysis), with values of 0.75, 0.49, 0.44, 0.54 and 0.67 – accuracy, F1, precision, recall and roc_auc, respectively.
- In the tests carried out in the second scenario of version A, there were 39 tests that obtained an f1 higher than the average f1 value for the scenario – 0.34. The model and experiments that obtained the best result (0.46) was LR with the experiments 2 and 3 – standard scaler and standard scaler plus principal component analysis – in which the list of features is *feat5*. With these experiences it achieved the following values 0.80, 0.46, 0.62, 0.37 for accuracy, precision, recall and roc_auc, respectively (Table 21). The worst experience was presented by the LR model with the same experiments as the best, however, the list of features used is different, using *feat2* and obtaining the values of 0.80, 0.30, 0.81, 0.19, 0.59 – accuracy, f1, precision, recall and roc_auc, respectively.
- Regarding scenario 2.b, there were 184 tests that obtained an f1 higher than the average f1 value for the scenario – 0.34. The model and experiment that obtained the best result (0.62) was MLP with experiment 4 – standard scaler plus PCA – using features list of feat5. With these experiences it achieved the following values 0.83, 0.61, 0.63, 0.76 for accuracy, precision, recall and ROC_AUC, respectively (Table 21). The worst result was achieved by the MLP model with experiment 4 – standard scaler with SMOTE – however, the list of features used is manual, with values of 0.75, 0.49, 0.44, 0.54 and 0.67 for accuracy, f1, Precision, Recall and ROC_AUC, respectively.
- In the tests carried out in the third scenario, there were 297 tests that obtained an f1 higher than the average f1 value for the scenario – 0.56. The model and experiment that obtained the best result (0.67) was RF with the standard scaler with SMOTE plus random under sampling experiment with the *feat7* feature list. With these experiences it achieved the following values 0.65, 0.73, 0.59, 0.71 for f1, precision, recall and ROC_AUC, respectively (Table 21). The worst

result was given by the RF model with the baseline experiment, however, the list of features used is *feat2*, with values of 0.67, 0.54, 0.69, 0.45, 0.64 – accuracy, f1, precision, recall and ROC_AUC, respectively.

Considering the experiences of scenarios 1, 2.b and 3 as the experiences with better performance, it was observed that withing the lists of features used (feat4 and feat5) there were several financial ratios calculated manually (as mentioned in Table 4):

- Growth rate net sales T3 (profitability)
- total assets to total liabilities (coverage)
- growth rate total assets t1 (efficiency)
- total debt / total assets (leverage)
- total assets
- operating cashflow current liabilities (profitability)
- inventory turnover (turnover)
- gross income divided by sales (profitability)
- growth rate net sales T2 (profitability)
- earnings before tax and interest/total asset (profitability)

Table 21. Metrics results from best overall experience from each scenario

Model	**Experience**	**Features list (number of features)**	**Accuracy**	**F1**	**Precision**	**Recall**	**ROC_AUC**	**Confusion Matrix**
Scenario 1								
SVC	4	feat5 (50)	0.72	0.66	0.63	0.70	0.72	TP:99 TN:160 FP:59 FN:42
SVC	5	feat5 (50)	0.72	0.66	0.63	0.70	0.72	TP:99 TN:160 FP:59 FN:42
Scenario 2.a								
LR	2	allFeat (267)	0.80	0.46	0.62	0.37	0.65	TP:26 TN:208 FP:16 FN:44
LR	3	allFeat (267)	0.80	0.46	0.62	0.37	0.65	TP:26 TN:208 FP:16 FN:44
Scenario 2.b								
MLP	5	feat5 (50)	0.83	0.62	0.61	0.63	0.76	TP:41 TN:203

								FP:26 FN:24
MLP	4	feat5 (50)	0.83	0.62	0.63	0.62	0.76	TP:40 TN:205 FP:24 FN:25
Scenario 3								
LR	4	feat4 (50)	0.71	0.67	0.66	0.68	0.70	TP:123 TN:164 FP:63 FN:57
LR	5	feat4 (50)	0.71	0.67	0.66	0.68	0.70	TP:123 TN:164 FP:63 FN:57

4.7. Improvements over previous scenarios

After analysing the results obtained from the various test scenarios according to the different characteristics defined in the data set, the results of the defined scenarios were improved. The improvement was not presented as another scenario because a scenario represents a data set with a defined characteristic, as shown in Table 12. Therefore, taking the best experiments extracted in scenarios 1, 2.b and 3, two more experiments were carried out to check their impact on the results: adding more features to the experiment's list of features and hyper parameterisation when training models. The scenario 2.a was not selected for improvement because the results presented by f1 (0.46), and recall (0.37) were not interesting.

In an attempt to see if there were any improvements by adding the features used in the *manualFeat* list (calculated financial ratios), which were not inherent from the provided files to the best experiments in each scenario, it was concluded that the results did not improve and had a negative impact on the metrics results. Scenario 2.b had a steeper drop in results compared to scenarios 1 and 3, which in turn had a drop of 0.03 and 0.04 percent in the f1 score, respectively (Table 24).

The initial tests employed the default parameters during the machine learning model training, as detailed in the Table 38 in annex. The objective of hyper parameterization is to meticulously choose the most effective hyperparameters for the machine learning model to achieve optimal results. To accomplish this goal, one of the techniques employed to discover the optimal hyperparameter combinations was *GridSearchCV*, a methodical search approach utilizing a grid of hyperparameters. When identifying the best-performing tests for each scenario, it was observed that the utilization of both the SMOTE (resampler) and random under sampling (under sampler) was a common characteristic among the top-performing tests. Consequently, the values of hyperparameter - *sampling strategy* - applied to these "balancing" techniques were as follows: auto, 0.5, 0.75. The Table 22 describes all the hyperparameters used for each model and the Table 23 describes the best hyperparameters selected for each model by *GridSearchCV.*

Table 22. Hyperparameters used in *GridSearchCV*

Scenario	**Model**	**Hyperparameters**
1	SVC	C : 0.1, 1, 10 Kernel : linear, rbf
2.b	MLP	Activation: relu, tanh Alpha: 0.0001, 0.001, 0.01
3	LR	C: 0.001, 0.01, 0.1, 1, 10 Penalty: l1, l2 Solver: liblinear, saga

Table 23. Best hyperparameters selected by *GridSearchCV*

Scenario	**Experience**	**Model**	**Hyperparameters**
1	4	SVC	classifier__C: 1, classifier__kernel: linear, resampler__sampling_strategy: auto
1	5	SVC	classifier__C: 10, classifier__kernel: rbf, resampler__sampling_strategy: auto, undersampler__sampling_strategy: auto
2.b	4	MLP	classifier__activation: tanh, classifier__alpha: 0.0001, resampler__sampling_strategy: auto
2.b	5	MLP	classifier__activation: tanh, classifier__alpha: 0.01, resampler__sampling_strategy: 0.75, undersampler__sampling_strategy: 0.5
3	4	LR	classifier__C: 0.1, classifier__penalty: l2, classifier__solver: liblinear, resampler__sampling_strategy: auto
3	5	LR	classifier__C: 0.001, classifier__penalty: l2, classifier__solver: saga, resampler__sampling_strategy: auto, undersampler__sampling_strategy: auto

With hyper parameterisation when training the models, it was found that the models in scenario 1 with experience 5 had an increase in the results of the f1 and precision metrics and the remaining metrics suffered a reduction in values. It is important to emphasise that experiment 5 of the SVC model in scenario 1 was the best experiment within the two improvement attempts, since it had the greatest increase in the f1 and precision metrics and the smallest decrease in results for accuracy, recall and roc_auc.

For scenario 2.b, it was notable the increase in f1 and recall precision. It was an increase of 27% for f1 (0.79) and 23.80% for recall (0.78), however, the accuracy score dropped to 0.78. In order to achieve a better capacity to distinguish between terminated projects from cancelled projects, this experience had a drop of accuracy result. With the hyper parameterisation, the model MLP proved to be the experiment with the best f1 result in all the tests carried out initially.

For scenario 3 with experience 4, there was a small evolution for the f1, precision and recall metrics, with the rise in the f1 metric being due to the rise in recall and precision. However, with the experience 5 the hyper parameterisation only dropped the results in comparison with the experience without it.

Table 24. Metrics results from improvements

Model (experience)	**Features list**	**Hyper parameterisation**	**Number of features**	**Accuracy**	**F1**	**Precision**	**Recall**	**ROC_AUC**
SVC (4)	feat5 + manual features		63	0.70	0.63	0.61	0.65	0.69
SVC (5)	feat5 + manual features		63	0.70	0.63	0.61	0.65	0.69
SVC (4)	feat5	X	50	0.64	0.65	0.66	0.64	0.64
SVC (5)	feat5	X	50	0.67	0.68	0.70	0.67	0.68
MLP (4)	feat5 + manual features		63	0.81	0.56	0.57	0.55	0.72
MLP (5)	feat5 + manual features		63	0.80	0.55	0.55	0.55	0.71
MLP (4)	feat5	X	50	0.78	0.79	0.79	0.78	0.70
MLP (5)	feat5	X	50	0.78	0.79	0.80	0.78	0.73

LR (4)	Feat4 + manual features		68	0.68	0.63	0.64	0.63	0.67
LR (5)	Feat4 + manual features		68	0.68	0.63	0.64	0.63	0.67
LR (4)	Feat4	X	50	0.70	0.70	0.70	0.70	0.69
LR (5)	Feat4	X	50	0.65	0.65	0.66	0.65	0.65

CHAPTER 5

Conclusion

5.1 Main conclusions

This dissertation is aimed at evaluating models capable of categorizing projects as either terminated or cancelled based on the information provided in the application as well as companies' financial ratios, utilizing machine learning models. It yielded pertinent findings regarding the application of artificial intelligence in project prediction, the primary determinants influencing classification, and the role of financial ratios in shaping the classification outcome.

The fact that the data set is imbalanced greatly affects the performance of the training model, and therefore experiments 4 (used synthetic minority oversampling technique – SMOTE) and 5 (used SMOTE and random under sampling) , in which synthetic data is generated for the class with the lowest frequency and under sampling (removal of records from the class with the highest number), can bring very interesting results, improving the results of the metrics used in this dissertation – accuracy, f1, precision, recall and roc_auc. With the best experiments carried out for the scenarios 1, 2.b and 3, it is possible to conclude:

- Although the accuracy presented by the scenarios 1 and 3 is lower than in scenario 2.a, the models presented by scenarios 1 and 3 perform better since the model has a better ability to classify completed or cancelled projects. The f1 metric values presented in scenarios 1 and 3 are noticeably higher than in scenario 2 (a and b), indicating that the models proposed in these scenarios are better at distinguishing completed projects (TP) from cancelled projects (TN). The reason for this behaviour is that in scenario 2.a, it was filtered out cancelled projects for which the reason was "promoter's withdrawal" and in scenario 2.b it was selected only cancelled projects which the reason was "promoter's withdrawal".
- Initially, the comparison between the experiments were made taking accuracy metric as main factor. However, the accuracy metric does not allow us to know if the model will perform well in classifying projects, so it is important to use other metrics such as f1 to see if the model classifies correctly. In scenario 2 (a and b), the best accuracy results were obtained, but it was not able to successfully distinguish between completed and cancelled projects. The fact that the cancelled projects were filtered out in more detail – by selecting (version a) or excluding (version b) only cancelled projects for which the reason was "Promoter's withdrawal" –, meant that the smallest class (cancelled projects) was left out, further highlighting the imbalance in the data set. In the beginning, the best f1 results from the scenario 2.b were lower in comparison to scenarios 1 and 3. However, after improving the model through hyper-

parameterisation, it was better at predicting completed and cancelled projects, resulting in a high f1 score, as can be seen in Table 24.

From the various tests carried out with different features, it can be concluded that the features based solely on financial ratio calculations (*manualFeat*) are not enough to obtain interesting results like the other features extracted through feature selection methods. Since the metrics and sampling results varied across the four test scenarios for experiments utilizing the manualFeat feature list, a test was conducted to assess how the list ratios affected project classification. This test involved adding features to the top-performing experiments extracted from scenarios 1, 2.b, and 3.The result was that adding these financial ratios had no positive impact on the trained models, leading to a slight degradation of the results, as shown in Table 24. It is important to note that the lists of features 4 and 5 – extracted by mutual information and decision tree, respectively – used in the best experiments obtained in scenarios 1, 2.b and 3 already had some financial ratios calculated manually as detailed in section 4.6 - *Best experience from each scenario*. Even using all the existing features in the dataset does not always guarantee the best performance as observed from Table 18.

With greater emphasis in experiences from scenarios 1 and 3, since the average values presented by the f1 metric – 0.67 and 0.65 – are higher than the average of scenarios 2.a and 2.b – 0.39 and 0.53, according to Tables 27, 29, 31, 33 in annex.

Therefore, it can be concluded that, the models and features that are more apt to help classify projects at the time of project application are those presented in scenarios 1, 2.b and 3. Models trained in scenario 2.a have a strong bias towards completed projects and perform poorly in predicting cancelled projects.

The best models presented in scenarios 1, 2.b and 3 are able to differentiate and classify completed projects from cancelled ones as seen in Table 21. As these three models have f1 values greater than 0.5, they are models that can decently predict both types of projects. The following list mentions the best model for each scenario and its characteristics:

- Scenario 1: Support Vector Classification (SVC) model with the experiment in which the features were pre-processed to the same scale (standard scaler) and the use of SMOTE, plus another test in which under sampling was added to the last experiment mentioned. Both with the features extracted from feature importance with Decision Tree (*feat5* list).
- Scenario 2.b: Multi-Layer Perceptron (MLP) model with same characteristics from scenario 1, mentioned above.
- Scenario 3: Logistic Regression (LR) model with the experiment in which the features were pre-processed to the same scale and the use of SMOTE, plus another test in which under sampling was added to the last experiment mentioned. The features used were extracted from mutual information (*feat4* list).

In order to see if it was possible to improve the results of the best performing experiments in each scenario - 1, 2.b, 3 - two separate methods were applied: adding manually calculated financial ratios to the experiment and tuning the training models.

Adding the calculated financial ratios to the list of the best features of each experiment is not enough to increase the metrics, it even had the opposite effect in some cases, producing slightly worse results. Therefore, it can be concluded that a greater number of features used for training is not equivalent to better results, the quality of the feature itself is important for classification problems and that with the features presented in the files it is possible to build models with the ability to predict cancelled projects from completed ones given the ratios presented in the application.

For the improvement experiment through hyper parameterisation, there were important improvements in metrics such as f1 and recall. For the models in scenarios 1, 2.b and 3, there were improvements in f1, but this compromised the accuracy result. However, for scenario 2.b, hyper parameterisation had very interesting results since the f1, precision and recall metrics had a huge increase in their results, while accuracy and roc_auc had a decrease of value. Concluding, that with the hyper parameterisation experiment there is an accuracy trade-off for increasing the f1 value, in some scenarios, indicating that the model has a slightly better capacity to distinguish between the 2 types of projects.

By concluding the project, it was possible to answer the questions posed in Chapter 1:

1. What conclusions and outcomes can be drawn from previous research that are relevant to the prediction of expense funding cancellation? From the previous studies (literature review), the financial ratios with the greatest impact on predicting a company's bankruptcy were identified and are described in Table 4.
2. Are financial ratios used in business failure useful to predict the project cancelation? In the stage of selecting the features with the greatest impact on the classification, some financial ratios were selected. However, according to the experiment in which the remaining financial ratios were added to the best experiment, the desired results were not obtained. It had a negative impact on the results, producing slightly worst results.
3. What are the ratios of a company that have a strong link to project cancellation? The manually calculated ratios, described in Table 4, which had the greatest impact on the project's cancellation classification are found in the lists of features of the best experiences for scenarios 1, 2.b and 3 in Table 21. Therefore, the lists of features selected by mutual information (*feat4*) and feature importance by decision tree (*feat5*) were the features of the best experiments, and within this list there were financial ratios calculated manually as detailed above in section 5.1.

4. Which information, besides financial ratios, have a strong impact on project cancellation? According to the features of the experience that achieved the best results (*feat4*), in addition to the features related to the financial ratios used, there is information available in the files that also impacts the classification of projects, such as: greater specificity of non-current assets (fixed assets), current assets (assets), and liabilities, as well as expenses before and after depreciation, team remuneration and building acquisition costs.
5. What is the performance of machine learning models on the prediction of project cancellation using the selected financial ratios? Of the various scenarios tested, the best experiences drawn from each scenario using only the financial ratios are described in the Table 25. Putting more emphasis on the f1 metric value, it concludes the experience using only financial ratios was most successful in scenario 3, yielding a result of 0.65. Another experiment with an interesting outcome occurred in scenario 1, achieving a result of 0.58. The difference between these two scenarios lies in the size and balance of the two classes in the dataset.

Table 25. Best experiences using only calculated financial ratios

Model	**Pre processing experience**	**Accuracy**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Scenario 1						
LSVC	5	0.64	0.59	0.53	0.66	0.64
Scenario 2.a						
SLP	2	0.79	0.43	0.59	0.34	0.63
SLP	3	0.79	0.43	0.59	0.34	0.63
Scenario 2.b						
MLP	4	0.75	0.49	0.44	0.54	0.67
Scenario 3						
RF	5	0.70	0.65	0.66	0.64	0.69

5.2 Contributions to the scientific and business community

With the work carried out throughout this dissertation, it will be possible to carry out a more thorough assessment of the project applications submitted to the lenders. This will allow for a faster and more effective assessment of applications, with the aim of predicting the projects with the lowest risk of

being cancelled and thus approving those applications. As a result, the lender will begin to accept projects more frequently and more quickly that are more likely to be successfully completed.

5.3 Study limitations

One of the primary challenges faced during the dissertation was locating articles that addressed the central research question. Nonetheless, efforts were made to explore articles related to financial ratios in the context of business banking and variables associated with project failure, aiming to approach the issue of project cancellations during the application phase as closely as possible. Using artificial intelligence approaches to examine the data presented in project applications, the research subject at hand is focused on the prediction of project cancellations. The prediction of company bankruptcies, where financial measures are used as predictive features, was found to be the issue most closely related to the existing body of literature.

Technical difficulties with this study project were mostly caused by the dataset's very small size and a clear class imbalance, which was seen in the dataset's underrepresentation of cancelled projects. The Synthetic Minority Over-sampling Technique (SMOTE) had to be used to artificially supplement the dataset with more examples in order to address the issue of class imbalance. It is important to recognise that it is difficult to draw broad-sweeping, universally applicable conclusions given the little amount of data that is currently accessible.

5.4. Future research proposals

For future research, we suggest acquiring a data set with a better balance between the classes so that the trained models have a good classification capacity.

Another suggestion would be to have access to the description and execution plan of the projects, and through text-mining techniques validate the possibility of cancelling the project given the plan in comparison with other successful projects.

The final suggestion would be to use data from macro-economic variables and verify that the success of a project is also due to the dependence of these factors.

Bibliographic references

[1] A. Fernández, “Artificial intelligence in financial services,” 2019.

[2] J. W. Wilcox, *A Simple Theory Of Financial Ratios As Predictors Of Failure*. 1970.

[3] M. Maricica and V. Georgeta, “Business Failure Risk Analysis using Financial Ratios,” *Procedia Soc Behav Sci*, vol. 62, pp. 728–732, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.123.

[4] R. O. Edmister, “Financial ratios as discriminant predictors of small business failure,” 1970.

[5] F. Ciampi and N. Gordini, “Using Economic-Financial Ratios for Small Enterprise Default Prediction Modeling: an Empirical Analysis,” University, St. Hugh’s College, 2008.

[6] A. R. Ahmad, Z. Azhar, and W. A. Wan-Abu-Bakar, “Cash-flows ratios as predictors of corporate failure,” in *ISIEA 2010 - 2010 IEEE Symposium on Industrial Electronics and Applications*, 2010, pp. 255–258. doi: 10.1109/ISIEA.2010.5679459.

[7] A. M. I. Lakshan, W. M. H. N. Wijekoon, and W. M. H. N. Wijekoon, “The Use of Financial Ratios in Predicting Corporate Failure in Sri Lanka Firm specific characteristics and voluntary disclosure View project The Use of Financial Ratios in Predicting Corporate Failure in Sri Lanka,” 2013, doi: 10.5176/2010-4804_2.4.249.

[8] F. Ciampi, A. Giannozzi, G. Marzi, and E. I. Altman, “Rethinking SME default prediction: a systematic literature review and future perspectives,” *Scientometrics*, vol. 126, no. 3, pp. 2141–2188, Mar. 2021, doi: 10.1007/s11192-020-03856-0.

[9] K. M. Poston, W. K. Harmon, and J. D. Gramlich, “A test of financial ratios as predictors of turnaround versus failure among financially distressed firms,” *Journal of Applied Business Research*, vol. 10, pp. 41–56.

[10] C. Y. Shirata, “Financial Ratios as Predictors of Bankruptcyin Japan: An Empirical Research,” 2012. [Online]. Available: https://www.researchgate.net/publication/228604054

[11] European Comission, “Annex II : Financial Guidelines.” [Online]. Available: http://ec.europa.eu/budget/inforeuro/index.cfm?Language=en

[12] P. Nový and I. M. Martišková, “An analysis of the determinants influencing the probability of cancellation for projects receiving EU funding,” 2013.

[13] A. L. Radu, E. A. Căldăraru, and M. Dimitriu, “Specific Features in Accessing European Funding.”

[14] H. Soroka-Potrzebna, “Risk identification as a basic stage of the project risk management,” *European Journal of Service Management*, vol. 27, pp. 277–283, 2018, doi: 10.18276/ejsm.2018.27/1-35.

[15] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable Machine Learning in Credit Risk Management,” *Comput Econ*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[16] S. Shi, R. Tse, W. Luo, S. D’Addona, and G. Pau, “Machine learning-driven credit risk: a systemic review,” *Neural Computing and Applications*, vol. 34, no. 17. Springer Science and Business Media Deutschland GmbH, pp. 14327–14339, Sep. 01, 2022. doi: 10.1007/s00521-022-07472-2.

[17] S. Kotsiantis, D. Tzelepis, S. Kotsiantis, E. Koumanakos, and V. Tampakas, “Efficiency of Machine Learning Techniques in Bankruptcy Prediction Ensemble methods View project Educational Data Mining View project Efficiency of Machine Learning Techniques in Bankruptcy Prediction,” 2005. [Online]. Available: https://www.researchgate.net/publication/215990968

[18] G.-D. Bordeianu and F. Radu, “Basic Types of Financial Ratios Used to Measure a Company’s Performance,” 2020. [Online]. Available: www.ugb.ro/etc

[19] A. Aziz and A. A. Rahman, “International Journal of Economics and Financial Issues The Relationship between Solvency Ratios and Profitability Ratios: Analytical Study in Food Industrial Companies listed in Amman Bursa,” *International Journal of Economics and Financial Issues*, vol. 7, no. 2, pp. 86–93, 2017, [Online]. Available: http:www.econjournals.com

[20] S. Nickolas, “What Do Efficiency Ratios Measure?,” https://www.investopedia.com/ask/answers/040715/what-do-efficiency-ratios-measure.asp.

[21] University of Maryland, “Financial Ratios & Formulas”.

[22] “Investopedia - dictionary.” Accessed: Jun. 04, 2023. [Online]. Available: https://www.investopedia.com/financial-term-dictionary-4769738

[23] Tim Stobierski, “How to Read & Understand a Balance Sheet,” https://online.hbs.edu/blog/post/how-to-read-a-balance-sheet.

[24] A. Yahya and S. Hidayat, “The Influence of Current Ratio, Total Debt to Total Assets, Total Assets Turn Over, and Return on Assets on Earnings Persistence in Automotive Companies,” *Journal of Accounting Auditing and Business*, vol. 3, no. 1, p. 62, Jan. 2020, doi: 10.24198/jaab.v3i1.24959.

[25] M. Ben, “Investopedia - How to Calculate an Altman Z-Score.” Accessed: Jun. 04, 2023. [Online]. Available: https://www.investopedia.com/articles/fundamental/04/021104.asp

[26] “Planning Analysis: Calculating Growth Rates,” https://pages.uoregon.edu/rgp/PPPM613/class8a.htm.

[27] J. Singh and P. Yadav, “RETURN ON CAPITAL EMPLOYED OF BANKING COMPANIES INCLUDED IN NIFTY: A STUDY,” 2017.

[28] Diné College, “References guide to basic accounting formulas,” References guide to basic accounting formulas.

[29] H. Agha, “IMPACT OF WORKING CAPITAL MANAGEMENT ON PROFITABILITY,” 2014.

Annexes and appendices

Annex A

Table 26. Table of f1 results by model and pre-processing experience

Model	Pre-processing experience	F1
Scenario 1		
GaussianNaiveBayes	1	0.22
GaussianNaiveBayes	2	0.35
GaussianNaiveBayes	3	0.52
GaussianNaiveBayes	4	0.49
GaussianNaiveBayes	5	0.49
K-NearestNeighbors	1	0.49
K-NearestNeighbors	2	0.56
K-NearestNeighbors	3	0.56
K-NearestNeighbors	4	0.6
K-NearestNeighbors	5	0.6
LinearSupportVectorClassification	2	0.56
LinearSupportVectorClassification	3	0.56
LinearSupportVectorClassification	4	0.61
LinearSupportVectorClassification	5	0.62
LogisticRegression	1	0.43
LogisticRegression	2	0.56
LogisticRegression	3	0.56
LogisticRegression	4	0.61
LogisticRegression	5	0.61
MLPClassifier	1	0.53
MLPClassifier	2	0.56
MLPClassifier	3	0.56
MLPClassifier	4	0.59
MLPClassifier	5	0.61
Perceptron	1	0.38
Perceptron	2	0.58
Perceptron	3	0.58
Perceptron	4	0.56

Perceptron	5	0.59
SupportVectorClassification	2	0.48
SupportVectorClassification	3	0.48
SupportVectorClassification	4	0.66
SupportVectorClassification	5	0.66
XGBoost	3	0.58
Scenario 2.a		
GaussianNaiveBayes	1	0.31
GaussianNaiveBayes	2	0.34
GaussianNaiveBayes	3	0.32
GaussianNaiveBayes	4	0.33
GaussianNaiveBayes	5	0.33
K-NearestNeighbors	2	0.43
K-NearestNeighbors	3	0.43
LinearSupportVectorClassification	2	0.4
LinearSupportVectorClassification	3	0.4
LogisticRegression	1	0.35
LogisticRegression	2	0.46
LogisticRegression	3	0.46
MLPClassifier	1	0.33
MLPClassifier	2	0.44
MLPClassifier	3	0.43
Perceptron	1	0.43
Perceptron	2	0.43
Perceptron	3	0.43
SupportVectorClassification	2	0.31
SupportVectorClassification	3	0.31
XGBoost	3	0.42
Scenario 2.b		
GaussianNaiveBayes	1	0.22
GaussianNaiveBayes	2	0.34
GaussianNaiveBayes	3	0.32
GaussianNaiveBayes	4	0.22

GaussianNaiveBayes	5	0.22
K-NearestNeighbors	1	0.36
K-NearestNeighbors	2	0.48
K-NearestNeighbors	3	0.48
LinearSupportVectorClassification	1	0.18
LinearSupportVectorClassification	2	0.34
LinearSupportVectorClassification	3	0.34
LogisticRegression	1	0.35
LogisticRegression	2	0.44
LogisticRegression	3	0.44
MLPClassifier	1	0.44
MLPClassifier	2	0.55
MLPClassifier	3	0.56
MLPClassifier	4	0.62
MLPClassifier	5	0.62
Perceptron	1	0.08
Perceptron	2	0.48
Perceptron	3	0.48
Perceptron	4	0.3
SupportVectorClassification	1	0
SupportVectorClassification	2	0.22
SupportVectorClassification	3	0.22
SupportVectorClassification	4	0.58
SupportVectorClassification	5	0.58
XGBoost	3	0.43
Scenario 3		
DecisionTree	1	0.63
DecisionTree	2	0.63
DecisionTree	3	0.61
DecisionTree	4	0.61
DecisionTree	5	0.61
GaussianNaiveBayes	1	0.16
GaussianNaiveBayes	2	0.55

GaussianNaiveBayes	3	0.6
GaussianNaiveBayes	4	0.63
GaussianNaiveBayes	5	0.63
K-NearestNeighbors	1	0.59
K-NearestNeighbors	2	0.65
K-NearestNeighbors	3	0.65
K-NearestNeighbors	4	0.65
K-NearestNeighbors	5	0.53
LinearSupportVectorClassification	1	0.62
LinearSupportVectorClassification	2	0.62
LinearSupportVectorClassification	3	0.62
LinearSupportVectorClassification	4	0.65
LinearSupportVectorClassification	5	0.65
LogisticRegression	1	0.53
LogisticRegression	2	0.61
LogisticRegression	3	0.61
LogisticRegression	4	0.67
LogisticRegression	5	0.67
MLPClassifier	1	0.59
MLPClassifier	2	0.63
MLPClassifier	3	0.64
MLPClassifier	4	0.66
MLPClassifier	5	0.66
Perceptron	1	0.63
Perceptron	2	0.65
Perceptron	3	0.65
Perceptron	4	0.67
Perceptron	5	0.67
RandomForest	1	0.64
RandomForest	2	0.64
RandomForest	3	0.61
RandomForest	4	0.65
RandomForest	5	0.65

SupportVectorClassification	1	0.64
SupportVectorClassification	2	0.55
SupportVectorClassification	3	0.55
SupportVectorClassification	4	0.65
SupportVectorClassification	5	0.65
XGBoost	1	0.65
XGBoost	2	0.65
XGBoost	3	0.62
XGBoost	4	0.65
XGBoost	5	0.65

Table 27. Results of tests from scenario 1 after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**Accuracy**	**F1**	**Precision**	**Recall**	**ROC_AUC**
SVC	4	feat5	50	0.72	0.66	0.63	0.70	0.72
SVC	5	feat5	50	0.72	0.66	0.63	0.70	0.72
SVC	4	feat4	50	0.70	0.62	0.61	0.63	0.69
SVC	5	feat4	50	0.70	0.62	0.61	0.63	0.69
LSVC	5	feat4	50	0.68	0.62	0.58	0.66	0.67
SVC	4	feat6	50	0.68	0.61	0.57	0.66	0.67
SVC	5	feat6	50	0.68	0.61	0.57	0.66	0.67
LSVC	4	feat4	50	0.68	0.61	0.58	0.65	0.67
MLP	5	feat2	30	0.68	0.61	0.58	0.65	0.67
LSVC	5	allFeat	267	0.69	0.61	0.59	0.62	0.68
LR	4	feat4	50	0.68	0.61	0.58	0.65	0.67
LR	5	feat4	50	0.68	0.61	0.58	0.65	0.67
SVC	4	feat3	50	0.69	0.60	0.61	0.60	0.68
SVC	5	feat3	50	0.69	0.60	0.61	0.60	0.68
LSVC	5	feat6	50	0.68	0.60	0.59	0.61	0.67
KNN	4	feat3	50	0.68	0.60	0.58	0.62	0.66
KNN	5	feat3	50	0.68	0.60	0.58	0.62	0.66
LSVC	4	allFeat	267	0.67	0.60	0.58	0.62	0.66
LSVC	4	feat5	50	0.68	0.60	0.58	0.61	0.66

MLP	4	feat2	30	0.67	0.59	0.57	0.62	0.66
LR	4	feat5	50	0.67	0.59	0.57	0.62	0.66
LR	5	feat5	50	0.67	0.59	0.57	0.62	0.66
LSVC	5	feat5	50	0.67	0.59	0.58	0.61	0.66
LSVC	4	feat6	50	0.68	0.59	0.59	0.60	0.66
MLP	5	feat7	50	0.69	0.59	0.61	0.57	0.67
LR	4	feat1	30	0.64	0.59	0.53	0.66	0.64
LR	5	feat1	30	0.64	0.59	0.53	0.66	0.64
SVC	4	allFeat	267	0.69	0.59	0.62	0.57	0.67
SVC	5	allFeat	267	0.69	0.59	0.62	0.57	0.67
SVC	4	feat7	50	0.66	0.59	0.56	0.62	0.66
SVC	5	feat7	50	0.66	0.59	0.56	0.62	0.66
SVC	4	feat2	30	0.67	0.59	0.57	0.61	0.66
SVC	5	feat2	30	0.67	0.59	0.57	0.61	0.66
LSVC	4	feat1	30	0.65	0.59	0.54	0.65	0.65
LSVC	5	feat1	30	0.65	0.59	0.54	0.65	0.65
LSVC	5	manualFeat	22	0.64	0.59	0.53	0.66	0.64
MLP	4	feat7	50	0.68	0.59	0.60	0.57	0.66
MLP	5	feat4	50	0.68	0.59	0.60	0.57	0.66
SLP	5	allFeat	267	0.69	0.59	0.62	0.55	0.67
LSVC	4	feat7	50	0.64	0.59	0.53	0.65	0.64
LSVC	5	feat7	50	0.64	0.59	0.53	0.65	0.64
LR	4	allFeat	267	0.67	0.58	0.57	0.60	0.65
LR	5	allFeat	267	0.67	0.58	0.57	0.60	0.65
LSVC	4	manualFeat	22	0.64	0.58	0.53	0.65	0.64
LSVC	5	feat3	50	0.64	0.58	0.54	0.62	0.64
XGBoost	3	feat1	30	0.68	0.58	0.60	0.56	0.66
MLP	4	feat6	50	0.68	0.58	0.59	0.57	0.66
LR	4	feat6	50	0.67	0.58	0.57	0.58	0.65
LR	5	feat6	50	0.67	0.58	0.57	0.58	0.65
SVC	4	manualFeat	22	0.66	0.58	0.56	0.60	0.65

Table 28. Confusion matrix from scenario 1 after filtering dummy and overfit experiences (top 50)

Model	Experience	Feature list	Number of features	TP	TN	FP	FN
SVC	4	feat5	50	99	160	59	42
SVC	5	feat5	50	99	160	59	42
SVC	4	feat4	50	89	163	56	52
SVC	5	feat4	50	89	163	56	52
LSVC	5	feat4	50	93	151	68	48
SVC	4	feat6	50	93	150	69	48
SVC	5	feat6	50	93	150	69	48
LSVC	4	feat4	50	92	151	68	49
MLP	5	feat2	30	91	153	66	50
LSVC	5	allFeat	267	88	159	60	53
LR	4	feat4	50	91	152	67	50
LR	5	feat4	50	91	152	67	50
SVC	4	feat3	50	84	166	53	57
SVC	5	feat3	50	84	166	53	57
LSVC	5	feat6	50	86	159	60	55
KNN	4	feat3	50	87	156	63	54
KNN	5	feat3	50	87	156	63	54
LSVC	4	allFeat	267	87	155	64	54
LSVC	4	feat5	50	86	157	62	55
MLP	4	feat2	30	88	152	67	53
LR	4	feat5	50	87	154	65	54
LR	5	feat5	50	87	154	65	54
LSVC	5	feat5	50	86	156	63	55
LSVC	4	feat6	50	84	160	59	57
MLP	5	feat7	50	81	167	52	60
LR	4	feat1	30	93	138	81	48
LR	5	feat1	30	93	138	81	48
SVC	4	allFeat	267	80	169	50	61
SVC	5	allFeat	267	80	169	50	61
SVC	4	feat7	50	87	152	67	54
SVC	5	feat7	50	87	152	67	54

SVC	4	feat2	30	86	154	65	55
SVC	5	feat2	30	86	154	65	55
LSVC	4	feat1	30	91	142	77	50
LSVC	5	feat1	30	91	142	77	50
LSVC	5	manualFeat	22	93	137	82	48
MLP	4	feat7	50	81	165	54	60
MLP	5	feat4	50	81	165	54	60
SLP	5	allFeat	267	78	172	47	63
LSVC	4	feat7	50	92	138	81	49
LSVC	5	feat7	50	92	138	81	49
LR	4	allFeat	267	84	156	63	57
LR	5	allFeat	267	84	156	63	57
LSVC	4	manualFeat	22	91	139	80	50
LSVC	5	feat3	50	88	144	75	53
XGBoost	3	feat1	30	79	166	53	62
MLP	4	feat6	50	80	163	56	61
LR	4	feat6	50	82	158	61	59
LR	5	feat6	50	82	158	61	59
SVC	4	manualFeat	22	84	153	66	57

Table 29. Results of tests from scenario 2.a after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**Accuracy**	**F1**	**Precision**	**Recall**
LR	2	allFeat	267	0.80	0.46	0.62	0.37
LR	3	allFeat	267	0.80	0.46	0.62	0.37
MLP	2	feat6	50	0.79	0.44	0.58	0.36
MLP	3	feat5	50	0.80	0.43	0.64	0.33
SLP	2	manualFeat	22	0.79	0.43	0.59	0.34
MLP	3	feat4	50	0.79	0.43	0.59	0.34
SLP	3	manualFeat	22	0.79	0.43	0.59	0.34

SLP	1	feat1	30	0.80	0.43	0.69	0.31
MLP	2	feat5	50	0.79	0.43	0.62	0.33
KNN	2	feat1	30	0.78	0.43	0.57	0.34
KNN	3	feat1	30	0.78	0.43	0.57	0.34
SLP	2	feat5	50	0.78	0.42	0.56	0.34
SLP	3	feat5	50	0.78	0.42	0.56	0.34
KNN	2	feat3	50	0.78	0.42	0.58	0.33
KNN	3	feat3	50	0.78	0.42	0.58	0.33
XGBoost	3	allFeat	267	0.80	0.42	0.68	0.30
XGBoost	3	feat6	50	0.80	0.42	0.68	0.30
MLP	2	feat4	50	0.79	0.42	0.61	0.31
LR	2	feat5	50	0.82	0.41	0.86	0.27
LR	3	feat5	50	0.82	0.41	0.86	0.27
KNN	2	feat2	30	0.81	0.40	0.76	0.27
KNN	3	feat2	30	0.81	0.40	0.76	0.27
KNN	2	allFeat	267	0.80	0.40	0.73	0.27
KNN	3	allFeat	267	0.80	0.40	0.73	0.27
LSVC	2	feat5	50	0.81	0.40	0.86	0.26
LSVC	3	feat5	50	0.81	0.40	0.86	0.26
XGBoost	3	feat2	30	0.81	0.39	0.82	0.26
XGBoost	3	feat1	30	0.79	0.39	0.61	0.29
KNN	2	feat5	50	0.80	0.38	0.75	0.26
KNN	3	feat5	50	0.80	0.38	0.75	0.26
MLP	3	feat2	30	0.80	0.37	0.77	0.24
MLP	2	feat2	30	0.80	0.36	0.80	0.23
MLP	2	feat1	30	0.80	0.35	0.76	0.23
LSVC	2	feat4	50	0.80	0.35	0.76	0.23
LR	2	feat4	50	0.80	0.35	0.76	0.23
LSVC	3	feat4	50	0.80	0.35	0.76	0.23
LR	3	feat4	50	0.80	0.35	0.76	0.23
LR	1	feat1	30	0.79	0.35	0.63	0.24
GNB	2	allFeat	267	0.81	0.34	0.88	0.21
LR	1	feat2	30	0.80	0.34	0.79	0.21

MLP	1	feat1	30	0.81	0.33	0.93	0.20
LR	2	feat6	50	0.79	0.33	0.68	0.21
LR	3	feat6	50	0.79	0.33	0.68	0.21
GNB	4	allFeat	267	0.80	0.33	0.88	0.20
GNB	5	allFeat	267	0.80	0.33	0.88	0.20
MLP	2	feat3	50	0.79	0.32	0.65	0.21
LSVC	2	feat6	50	0.79	0.32	0.65	0.21
GNB	3	feat1	30	0.79	0.32	0.65	0.21
GNB	3	feat2	30	0.79	0.32	0.65	0.21
LSVC	3	feat6	50	0.79	0.32	0.65	0.21

Table 30. Confusion matrix of tests from scenario 2.a after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**TP**	**TN**	**FP**	**FN**
LR	2	allFeat	267	26	208	16	44
LR	3	allFeat	267	26	208	16	44
MLP	2	feat6	50	25	206	18	45
MLP	3	feat5	50	23	211	13	47
SLP	2	manualFeat	22	24	207	17	46
MLP	3	feat4	50	24	207	17	46
SLP	3	manualFeat	22	24	207	17	46
SLP	1	feat1	30	22	214	10	48
MLP	2	feat5	50	23	210	14	47
KNN	2	feat1	30	24	206	18	46
KNN	3	feat1	30	24	206	18	46
SLP	2	feat5	50	24	205	19	46
SLP	3	feat5	50	24	205	19	46
KNN	2	feat3	50	23	207	17	47
KNN	3	feat3	50	23	207	17	47

XGBoost	3	allFeat	267	21	214	10	49
XGBoost	3	feat6	50	21	214	10	49
MLP	2	feat4	50	22	210	14	48
LR	2	feat5	50	19	221	3	51
LR	3	feat5	50	19	221	3	51
KNN	2	feat2	30	19	218	6	51
KNN	3	feat2	30	19	218	6	51
KNN	2	allFeat	267	19	217	7	51
KNN	3	allFeat	267	19	217	7	51
LSVC	2	feat5	50	18	221	3	52
LSVC	3	feat5	50	18	221	3	52
XGBoost	3	feat2	30	18	220	4	52
XGBoost	3	feat1	30	20	211	13	50
KNN	2	feat5	50	18	218	6	52
KNN	3	feat5	50	18	218	6	52
MLP	3	feat2	30	17	219	5	53
MLP	2	feat2	30	16	220	4	54
MLP	2	feat1	30	16	219	5	54
LSVC	2	feat4	50	16	219	5	54
LR	2	feat4	50	16	219	5	54
LSVC	3	feat4	50	16	219	5	54
LR	3	feat4	50	16	219	5	54
LR	1	feat1	30	17	214	10	53
GNB	2	allFeat	267	15	222	2	55
LR	1	feat2	30	15	220	4	55
MLP	1	feat1	30	14	223	1	56
LR	2	feat6	50	15	217	7	55
LR	3	feat6	50	15	217	7	55
GNB	4	allFeat	267	14	222	2	56
GNB	5	allFeat	267	14	222	2	56
MLP	2	feat3	50	15	216	8	55
LSVC	2	feat6	50	15	216	8	55
GNB	3	feat1	30	15	216	8	55

GNB	3	feat2	30	15	216	8	55
LSVC	3	feat6	50	15	216	8	55

Table 31. Results of tests from scenario 2.b after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**Accuracy**	**F1**	**Precision**	**Recall**
MLP	5	feat5	50	0.83	0.62	0.61	0.63
MLP	4	feat5	50	0.83	0.62	0.63	0.62
SVC	4	feat5	50	0.79	0.58	0.51	0.66
SVC	5	feat5	50	0.79	0.58	0.51	0.66
LSVC	4	feat7	50	0.74	0.57	0.45	0.77
SVC	4	feat6	50	0.76	0.57	0.47	0.72
SVC	5	feat6	50	0.76	0.57	0.47	0.72
LSVC	5	feat7	50	0.74	0.57	0.45	0.77
MLP	3	feat5	50	0.84	0.56	0.69	0.48
LR	4	feat7	50	0.74	0.56	0.45	0.75
LR	5	feat7	50	0.74	0.56	0.45	0.75
LR	4	feat5	50	0.75	0.55	0.46	0.71
LR	5	feat5	50	0.75	0.55	0.46	0.71
LSVC	5	feat3	50	0.75	0.55	0.45	0.69
MLP	2	feat5	50	0.84	0.55	0.71	0.45
LSVC	4	feat3	50	0.74	0.54	0.44	0.69
LSVC	5	feat4	50	0.70	0.52	0.41	0.74
LSVC	5	feat5	50	0.73	0.52	0.43	0.66
LSVC	4	feat4	50	0.70	0.52	0.40	0.74
MLP	3	feat6	50	0.81	0.52	0.60	0.46
LSVC	4	feat5	50	0.73	0.52	0.43	0.66
LR	4	feat3	50	0.72	0.52	0.42	0.68
LR	5	feat3	50	0.72	0.52	0.42	0.68
LR	4	feat4	50	0.70	0.52	0.40	0.74
LR	5	feat4	50	0.70	0.52	0.40	0.74
MLP	2	feat7	50	0.82	0.52	0.65	0.43

SVC	4	feat7	50	0.73	0.52	0.43	0.65
SVC	5	feat7	50	0.73	0.52	0.43	0.65
MLP	4	feat6	50	0.78	0.52	0.51	0.52
SLP	5	feat4	50	0.67	0.51	0.38	0.80
LR	4	feat6	50	0.71	0.51	0.41	0.69
LR	5	feat6	50	0.71	0.51	0.41	0.69
LSVC	5	feat1	30	0.70	0.51	0.40	0.72
LSVC	4	feat6	50	0.71	0.51	0.41	0.69
LSVC	5	feat6	50	0.71	0.51	0.41	0.69
MLP	5	feat6	50	0.78	0.51	0.50	0.52
LSVC	4	feat1	30	0.69	0.51	0.39	0.72
LR	4	allFeat	267	0.76	0.51	0.46	0.57
SVC	4	allFeat	267	0.76	0.51	0.46	0.57
LR	5	allFeat	267	0.76	0.51	0.46	0.57
SVC	5	allFeat	267	0.76	0.51	0.46	0.57
SVC	4	feat4	50	0.71	0.51	0.40	0.68
SVC	5	feat4	50	0.71	0.51	0.40	0.68
SLP	2	feat2	30	0.68	0.51	0.38	0.74
SLP	3	feat2	30	0.68	0.51	0.38	0.74
SLP	5	feat1	30	0.62	0.50	0.35	0.88
MLP	4	feat7	50	0.78	0.50	0.50	0.51
LR	4	feat1	30	0.69	0.50	0.39	0.71
LR	5	feat1	30	0.69	0.50	0.39	0.71
SLP	5	feat3	50	0.60	0.49	0.34	0.88

Table 32. Confusion matrix of tests from scenario 2.b after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**TP**	**TN**	**FP**	**FN**
MLP	5	feat5	50	41	203	26	24
MLP	4	feat5	50	40	205	24	25
SVC	4	feat5	50	43	188	41	22

SVC	5	feat5	50	43	188	41	22
LSVC	4	feat7	50	50	169	60	15
SVC	4	feat6	50	47	176	53	18
SVC	5	feat6	50	47	176	53	18
LSVC	5	feat7	50	50	168	61	15
MLP	3	feat5	50	31	215	14	34
LR	4	feat7	50	49	169	60	16
LR	5	feat7	50	49	169	60	16
LR	4	feat5	50	46	174	55	19
LR	5	feat5	50	46	174	55	19
LSVC	5	feat3	50	45	175	54	20
MLP	2	feat5	50	29	217	12	36
LSVC	4	feat3	50	45	172	57	20
LSVC	5	feat4	50	48	159	70	17
LSVC	5	feat5	50	43	173	56	22
LSVC	4	feat4	50	48	158	71	17
MLP	3	feat6	50	30	209	20	35
LSVC	4	feat5	50	43	172	57	22
LR	4	feat3	50	44	169	60	21
LR	5	feat3	50	44	169	60	21
LR	4	feat4	50	48	157	72	17
LR	5	feat4	50	48	157	72	17
MLP	2	feat7	50	28	214	15	37
SVC	4	feat7	50	42	174	55	23
SVC	5	feat7	50	42	174	55	23
MLP	4	feat6	50	34	196	33	31
SLP	5	feat4	50	52	144	85	13
LR	4	feat6	50	45	164	65	20
LR	5	feat6	50	45	164	65	20
LSVC	5	feat1	30	47	158	71	18
LSVC	4	feat6	50	45	163	66	20
LSVC	5	feat6	50	45	163	66	20
MLP	5	feat6	50	34	195	34	31

LSVC	4	feat1	30	47	157	72	18
LR	4	allFeat	267	37	185	44	28
SVC	4	allFeat	267	37	185	44	28
LR	5	allFeat	267	37	185	44	28
SVC	5	allFeat	267	37	185	44	28
SVC	4	feat4	50	44	164	65	21
SVC	5	feat4	50	44	164	65	21
SLP	2	feat2	30	48	152	77	17
SLP	3	feat2	30	48	152	77	17
SLP	5	feat1	30	57	125	104	8
MLP	4	feat7	50	33	196	33	32
LR	4	feat1	30	46	157	72	19
LR	5	feat1	30	46	157	72	19
SLP	5	feat3	50	57	119	110	8

Table 33. Results of tests from scenario 3 after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**Accuracy**	**F1**	**Precision**	**Recall**
LR	4	feat4	50	0.71	0.67	0.66	0.68
LR	5	feat4	50	0.71	0.67	0.66	0.68
SLP	4	feat4	50	0.68	0.67	0.61	0.74
SLP	5	feat4	50	0.67	0.67	0.60	0.75
MLP	4	feat2	30	0.70	0.66	0.66	0.66
MLP	5	feat2	30	0.70	0.66	0.66	0.66
MLP	5	feat4	50	0.70	0.66	0.66	0.66
MLP	5	feat6	50	0.70	0.66	0.65	0.66
XGBoost	1	allFeat	267	0.72	0.65	0.72	0.60
XGBoost	2	allFeat	267	0.72	0.65	0.72	0.60
MLP	5	feat3	50	0.69	0.65	0.64	0.67
SLP	2	allFeat	267	0.65	0.65	0.58	0.74
SLP	3	allFeat	267	0.65	0.65	0.58	0.74

RF	5	manualFeat	22	0.70	0.65	0.66	0.64
RF	5	feat7	50	0.72	0.65	0.73	0.59
LR	4	feat2	30	0.65	0.65	0.58	0.74
LR	5	feat2	30	0.65	0.65	0.58	0.74
KNN	2	feat3	50	0.71	0.65	0.70	0.61
KNN	3	feat3	50	0.71	0.65	0.70	0.61
KNN	4	feat3	50	0.68	0.65	0.63	0.67
LSVC	4	feat2	30	0.65	0.65	0.58	0.73
RF	4	feat7	50	0.72	0.65	0.73	0.58
MLP	4	feat6	50	0.70	0.65	0.66	0.64
RF	4	feat6	50	0.71	0.65	0.69	0.62
LSVC	5	feat2	30	0.65	0.65	0.58	0.73
KNN	4	feat4	50	0.69	0.65	0.65	0.64
MLP	5	feat7	50	0.69	0.65	0.65	0.64
SLP	2	feat1	30	0.69	0.65	0.66	0.64
SLP	3	feat1	30	0.69	0.65	0.66	0.64
SVC	4	feat4	50	0.70	0.65	0.67	0.63
SVC	5	feat4	50	0.70	0.65	0.67	0.63
RF	5	feat6	50	0.70	0.65	0.67	0.63
LSVC	5	allFeat	267	0.67	0.65	0.62	0.68
SLP	5	feat1	30	0.62	0.65	0.55	0.78
SLP	5	feat3	50	0.62	0.65	0.55	0.78
XGBoost	4	feat2	30	0.69	0.65	0.65	0.64
LSVC	4	feat3	50	0.69	0.65	0.65	0.64
MLP	4	feat3	50	0.69	0.65	0.65	0.64
MLP	4	feat5	50	0.69	0.65	0.65	0.64
XGBoost	5	feat2	30	0.69	0.65	0.65	0.64
LSVC	4	feat4	50	0.68	0.64	0.63	0.66
LSVC	5	feat4	50	0.68	0.64	0.63	0.66
LR	4	feat7	50	0.67	0.64	0.61	0.68
LR	5	feat7	50	0.67	0.64	0.61	0.68
SLP	4	feat5	50	0.62	0.64	0.55	0.78

RF	5	feat5	50	0.71	0.64	0.69	0.60
MLP	4	feat1	30	0.69	0.64	0.65	0.64
LSVC	5	feat7	50	0.66	0.64	0.61	0.68
XGBoost	4	feat6	50	0.69	0.64	0.65	0.63
XGBoost	5	feat6	50	0.69	0.64	0.65	0.63

Table 34. Confusion matrix of tests from scenario 3 after filtering dummy and overfit experiences (top 50)

Model	**Experience**	**Feature list**	**Number of features**	**TP**	**TN**	**FP**	**FN**
LR	4	feat4	50	123	164	63	57
LR	5	feat4	50	123	164	63	57
SLP	4	feat4	50	133	143	84	47
SLP	5	feat4	50	135	137	90	45
MLP	4	feat2	30	118	167	60	62
MLP	5	feat2	30	118	167	60	62
MLP	5	feat4	50	118	167	60	62
MLP	5	feat6	50	119	164	63	61
XGBoost	1	allFeat	267	108	185	42	72
XGBoost	2	allFeat	267	108	185	42	72
MLP	5	feat3	50	121	158	69	59
SLP	2	allFeat	267	134	131	96	46
SLP	3	allFeat	267	134	131	96	46
RF	5	manualFeat	22	116	168	59	64
RF	5	feat7	50	106	188	39	74
LR	4	feat2	30	133	132	95	47
LR	5	feat2	30	133	132	95	47
KNN	2	feat3	50	110	179	48	70
KNN	3	feat3	50	110	179	48	70
KNN	4	feat3	50	121	156	71	59
LSVC	4	feat2	30	132	133	94	48

RF	4	feat7	50	105	189	38	75
MLP	4	feat6	50	115	168	59	65
RF	4	feat6	50	111	176	51	69
LSVC	5	feat2	30	132	132	95	48
KNN	4	feat4	50	116	165	62	64
MLP	5	feat7	50	116	165	62	64
SLP	2	feat1	30	115	167	60	65
SLP	3	feat1	30	115	167	60	65
SVC	4	feat4	50	113	171	56	67
SVC	5	feat4	50	113	171	56	67
RF	5	feat6	50	113	171	56	67
LSVC	5	allFeat	267	122	152	75	58
SLP	5	feat1	30	140	114	113	40
SLP	5	feat3	50	140	114	113	40
XGBoost	4	feat2	30	115	166	61	65
LSVC	4	feat3	50	115	166	61	65
MLP	4	feat3	50	115	166	61	65
MLP	4	feat5	50	115	166	61	65
XGBoost	5	feat2	30	115	166	61	65
LSVC	4	feat4	50	118	159	68	62
LSVC	5	feat4	50	118	159	68	62
LR	4	feat7	50	123	148	79	57
LR	5	feat7	50	123	148	79	57
SLP	4	feat5	50	140	112	115	40
RF	5	feat5	50	108	179	48	72
MLP	4	feat1	30	115	164	63	65
LSVC	5	feat7	50	123	147	80	57
XGBoost	4	feat6	50	114	166	61	66
XGBoost	5	feat6	50	114	166	61	66

Table 35. Table of lists of features used in models training

List name	Number of features	Features
feat1	30	RES_ANTES_DEPRECIACAO_GASTOS, GASTOS_DEPRECIACAO_AMORTIZA, ATIVO_TOTAL, CP_E_PASSIVO_TOTAL, 2_PESSOAL_NHT_PSETC_REMUNERADAS, 1_PESSOAL_NMP_PSE_HOMENS, 2_PESSOAL_NHT_PSE_HOMENS, IAE_GASTOS_PESS_REMUN, Resumo/Dimensao, Dimensao/Dimensao, Analisemercados/Direcao, Header/Informa_Db, Txtfinanc/Fonte, Incentivo/Elegivel, Incentivo/Dimensao, Incentivo/Pt_Qualif, Incentivo/Limite_Pt_Qualif, Incentivo/Aut_Gestao, Incentivo/Tx_Limite, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Ii, Incentivo/Tx_Major_Eleg, Incentivo/Tx_Major_Eleg_Bd, Incentivo/Incentivo, Incentivo/Incentivo_Nr, EBITDA, EBIT, Total_Assets, Growth_Rate_Total_Assets_T3
feat2	30	GASTOS_PESSOAL, OUTROS_REDIMENTOS_GANHOS, RES_ANTES_IMPOSTOS, ATIVO_NCOR_TOTAL, ATIVO_COR_OUTRA_CONT_A_RECEBER, CP_TOTAL, PASSIVO_COR_OUT_CONTAS_A_PAGAR, 2_PESSOAL_NHT_PSERNR, 2_PESSOAL_NHT_PESS_REMUN_SE, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, 1_PESSOAL_NMP_PSETC_REMUNERADAS, 2_PESSOAL_NHT_PSETC_REMUNERADAS, GASTOS_PESSOAL_TOTAL, GP_ENCARG_REMUN, IAE_VENDAS, IAE_CMVMC_MATER_PRIMAS, IAE_NUM_MED_PESS_SERV, IAE_GASTOS_PESS, Resumo/Concelho, Header/Informa_Db, Incentivo/Pt_Qualif, Incentivo/Limite_Pt_Qualif, Incentivo/Tx_Limite, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Eleg, Incentivo/Tx_Major_Eleg_Bd, Incentivo/Incentivo, Incentivo/Incentivo_Nr, Growth_Rate_Net_Sales_T3
feat3	50	GASTOS_PESSOAL, RES_ANTES_DEPRECIACAO_GASTOS, GASTOS_DEPRECIACAO_AMORTIZA, RESULTADO_OPERACIONAL, ATIVO_COR_TOTAL, ATIVO_TOTAL, CP_TOTAL, CP_E_PASSIVO_TOTAL, 1_PESSOAL_NMP_PSERNR, 2_PESSOAL_NHT_PSERNR, 1_PESSOAL_NMP_PESS_REMUN_SE, 2_PESSOAL_NHT_PESS_REMUN_SE, 1_PESSOAL_NMP_PSE_TEMPO_COMPLETO, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, 1_PESSOAL_NMP_PSETC_REMUNERADAS, 2_PESSOAL_NHT_PSETC_REMUNERADAS, 1_PESSOAL_NMP_PSE_HOMENS, 2_PESSOAL_NHT_PSE_HOMENS, GASTOS_PESSOAL_TOTAL, GP_REMUN_PESSOAL, GP_ENCARG_REMUN, IAE_VENDAS_PAISDRR, IAE_NUM_MED_PESS_SERV, IAE_GASTOS_PESS, IAE_GASTOS_PESS_REMUN, 4_IMG_TOT_AQUIS_ACT_FIX_TANG, Resumo/Dimensao, Resumo/Nute_Lisboa, Dimensao/Dimensao, Analisemercados/Direcao, Header/Informa_Db, Txtfinanc/Fonte, Incentivo/Elegivel, Incentivo/Dimensao, Incentivo/Pt_Qualif, Incentivo/Limite_Pt_Qualif, Incentivo/Aut_Gestao,

		Incentivo/Tx_Limite, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Ii, Incentivo/Tx_Major_Eleg, Incentivo/Tx_Major_Eleg_Bd, Incentivo/Incentivo, Incentivo/Incentivo_Nr, Inovtecn/Racio_Dr, EBITDA, EBIT, Total_Assets, Growth_Rate_Total_Assets_T3
feat4	50	VENDAS_SERVICOS_PRESTADOS, GASTOS_PESSOAL, OUTROS_REDIMENTOS_GANHOS, RES_ANTES_DEPRECIACAO_GASTOS, RES_ANTES_IMPOSTOS, ATIVO_NCOR_TOTAL, CP_OUTRAS_RESERVAS, CP_OUTRAS_VARIACAOES_CAP_PRO, CP_RESULTADO_LIQUIDO_PERIODO, CP_TOTAL, PASSIVO_COR_OUT_CONTAS_A_PAGAR, PASSIVO_TOTAL, 2_PESSOAL_NHT_PSERNR, 2_PESSOAL_NHT_PESS_REMUN_SE, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, 1_PESSOAL_NMP_PSETC_REMUNERADAS, 2_PESSOAL_NHT_PSETC_REMUNERADAS, 1_PESSOAL_NMP_PSE_MULHERES, 2_PESSOAL_NHT_PSE_MULHERES, GASTOS_PESSOAL_TOTAL, GP_REMUN_ORGAOS_SOCIAIS, GP_ENCARG_REMUN, IAE_VENDAS_PAISDRR, IAE_CMVMC_MATER_PRIMAS, IAE_GASTOS_PESS, 4_IMG_TOT_VENDAS, 1_IMG_INT_AQUIS_ACT_FIX_TANG, 1_IMG_INT_REND_SUPLEM, Resumo/Concelho, Resumo/Dimensao, Resumo/Nute_Norte, Resumo/N_Pt_Pos, Promotor/Distrito, Promotor/Concelho, Txtfinanc/Fonte, Incentivo/Dimensao, Incentivo/Pt_Qualif, Incentivo/Aut_Gestao, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Ii, Incentivo/Base_Eleg, Incentivo/Tx_Major_Eleg, Incentivo/Tx_Major_Eleg_Bd, Incentivo/Incentivo, Incentivo/Incentivo_Nr, total_assets_to_total_liabilities, total debt / total assets, Growth_Rate_Net_Sales_T3, Growth_Rate_Total_Assets_T1
feat5	50	GASTOS_DEPRECIACAO_AMORTIZA, Incentivo/Tx_Limite, Operating cash flow current liabilities, Dadosprojecto/N_Meses, SUBSIDIOS_EXPLORACAO, ATIVO_COR_ESTADO_OUT_ENTES_PUB, PASSIVO_COR_OUT_CONTAS_A_PAGAR, GP_REMUN_ORGAOS_SOCIAIS, IAE_VAR_INVENT_PROD, ATIVO_COR_DIFERIMENTOS, Resumo/Cae, IAE_VENDAS_MERCADORIAS, 2_IMG_COM_VENDAS, total debt / total assets, PASSIVO_NC_FINANCIAMENTOS_OBTD, ATIVO_NCOR_INV_FINANC_PQ_ENTID, Resumo/Nute_Norte, 3_IMG_EC_COMPRAS, Growth_Rate_Net_Sales_T3, NIF_Prom_anon, GP_SEG_ACID_TRAB_DOEN_PROF, IAE_AFT_QUANT_ESCR_LIQ_FIN, earnings before tax and interest / total asset, ATIVO_NCOR_PART_FINAN_EQV_PAT, CP_OUTRAS_VARIACAOES_CAP_PRO, Resumo/Investimento, IAE_AFT_TOTAL_AQUIS_EDIF, CP_RESERVAS_LEGAIS, Growth_Rate_Net_Sales_T2, 1_IMG_INT_FORN_SEREXTERN, PASSIVO_COR_ESTADO_OUT_ENT_PUB, IAE_PREST_SERV, IMPOSTO_RENDIMENTO_PERIODO, ATIVO_COR_ACCIONISTAS_SOCIOS, Inventory_Turnover, Total_Assets, ATIVO_COR_OUTRA_CONT_A_RECEBER, Gross income divided by sales, PASSIVO_COR_FORNCEDORES, Resumo/Distrito, ATIVO_COR_INVENTARIOS, RES_ANTES_DEPRECIACAO_GASTOS,

		CP_E_PASSIVO_TOTAL, 2_PESSOAL_NHT_PSE_MULHERES, total_assets_to_total_liabilities, PASSIVO_COR_OUTROS_PAS_CORRENTES, IAE_CMVMC_MERCADORIAS, IAE_COMPRAS, Resumo/Elegivel, 3_IMG_EC_REND_SUPLEM
feat6	50	RES_ANTES_DEPRECIACAO_GASTOS, GASTOS_DEPRECIACAO_AMORTIZA, ATIVO_NCOR_FIXOS_TANGIVEIS, IAE_AFT_QUANT_ESCR_LIQ_FIN, CP_TOTAL, Operating cash flow current liabilities, earnings before tax and interest / total asset, 4_IMG_TOT_AQUIS_ACT_FIX_TANG, ATIVO_COR_CAIXA_DEP_BANCARIOS, Growth_Rate_Net_Sales_T1, Incentivo/Pt_Qualif, CP_SOMA, ATIVO_NCOR_TOTAL, NIF_Prom_anon, Resumo/Cae, Gross income divided by sales, PASSIVO_COR_ESTADO_OUT_ENT_PUB, Growth_Rate_Total_Assets_T1, Resumo/Elegivel, Dadosprojecto/Investimento, IAE_AFT_TOTAL_AQUIS, total_assets_to_total_liabilities, total debt / total assets, 2_PESSOAL_NHT_PSERNR, Total_Assets, IAE_NUM_MED_PESS_SERV, Promotor/Concelho, Growth_Rate_Net_Sales_T3, GP_ENCARG_REMUN, Dadosprojecto/N_Meses, OUTROS_GASTOS_PERDAS, PASSIVO_TOTAL, EBIT, PASSIVO_COR_OUT_CONTAS_A_PAGAR, OUTROS_REDIMENTOS_GANHOS, CP_RESULTADO_LIQUIDO_PERIODO, Growth_Rate_Total_Assets_T3, Resumo/Investimento, EBITDA, CP_RESULTADOS_TRANSITADOS, IAE_CMVMC_MATER_PRIMAS, Dadosprojecto/Elegivel, Average_Collection_Period_For_Receivables, 2_PESSOAL_NHT_PSE_HOMENS, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, Incentivo/Dimensao, ATIVO_COR_DIFERIMENTOS, 1_IMG_INT_AQUIS_ACT_FIX_TANG, GASTOS_PESSOAL, ATIVO_COR_ESTADO_OUT_ENTES_PUB
feat7	50	Txtfinanc/Fonte, Promotor/Nat_Jur, 3_IMG_EC_PREST_SERV, GASTOS_DEPRECIACAO_AMORTIZA, RES_ANTES_DEPRECIACAO_GASTOS, ATIVO_NCOR_PART_FINAN_EQV_PAT, Growth_Rate_Net_Sales_T3, ATIVO_NCOR_FIXOS_TANGIVEIS, ATIVO_NCOR_INV_FINANC_PQ_ENTID, Resumo/Lst_Po, 1_IMG_INT_AQUIS_ACT_INTANG, PROVISOES, CP_AJUST_EM_ACT_FINANCEIROS, PASSIVO_NC_FINANCIAMENTOS_OBTD, ATIVO_COR_INVENTARIOS, ATIVO_COR_ACCIONISTAS_SOCIOS, 3_IMG_EC_AQUIS_ACT_FIX_TANG, 2_N_PESSOAL_NHT_PSETP_REMUNERADAS, Resumo/Nute_Norte, 2_IMG_COM_AQUIS_ACT_FIX_TANG, 2_PESSOAL_NHT_PSE_TEMPO_PARCIAL, Critselb1/N_Mercados, Growth_Rate_Total_Assets_T3, CP_TOTAL, GP_REMUN_ORGAOS_SOCIAIS, 3_IMG_EC_COMPRAS, Growth_Rate_Net_Sales_T2, Growth_Rate_Net_Sales_T1, Operating cash flow current liabilities, IAE_CMVMC_MATER_PRIMAS, Resumo/Investimento, 1_PESSOAL_NMP_PSE_MULHERES, ATIVO_NCOR_PROPRI_INVESTIMENTO, ATIVO_COR_CLIENTES, Analisemercados/Direcao, 2_IMG_COM_FORN_SEREXTERN, Growth_Rate_Total_Assets_T2,

		ATIVO_COR_ADIANTAMENTOS_FORNEC, PASSIVO_COR_ADIANTA_DE_CLIENTES, IAE_AFT_TOTAL_AQUIS_EDIF, total_assets_to_total_liabilities, ATIVO_COR_ESTADO_OUT_ENTES_PUB, CMVMC, 1_IMG_INT_RS_OUTROS, CP_OUTRAS_RESERVAS, 1_PESSOAL_NMP_PSE_HOMENS, EBIT, IAE_AFT_TOTAL_AQUIS, Promotor/Concelho, CP_OUTRAS_VARIACAOES_CAP_PRO
allFeat	267	VENDAS_SERVICOS_PRESTADOS, SUBSIDIOS_EXPLORACAO, GANHOS_PERDAS_SUBSIDIARIAS, VARIA_INVENTARIOS_PRODUCAO, TRABALHOS_PROPRIA_ENTIDADE, CMVMC, FSE, GASTOS_PESSOAL, IMPARIDADE_INVENTARIOS, IMPARIDADE_DIVIDAS_A_RECEBER, PROVISOES, IMPARIDADE_INVENTA_N_AMORT_INVESTIMENTOS_N_DEPRECIAVEIS, OUTRAS_IMPARIDADES, AUMENTOS_RED_JUSTO_VALOR, OUTROS_REDIMENTOS_GANHOS, OUTROS_GASTOS_PERDAS, RES_ANTES_DEPRECIACAO_GASTOS, GASTOS_DEPRECIACAO_AMORTIZA, RESULTADO_OPERACIONAL, JUROS_RENDIME_SIMILARES_OBT, JUROS_GAST_SIMILARES_SUPORT, RES_ANTES_IMPOSTOS, IMPOSTO_RENDIMENTO_PERIODO, RESULTADO_LIQUIDO_PERIODO, ATIVO_NCOR_FIXOS_TANGIVEIS, ATIVO_NCOR_PROPRI_INVESTIMENTO, ATIVO_NCOR_GOODWILL, ATIVO_NCOR_INTANGIVEIS, ATIVO_NCOR_PART_FINAN_EQV_PAT, ATIVO_NCOR_PART_FIN_OUTROS_MET, ATIVO_NCOR_ACCIONISTAS_SOCIOS, ATIVO_NCOR_OUTROS_ACT_FINANC, ATIVO_NCOR_IMPOSTOS_DIFERIDOS, ATIVO_NCOR_INV_FINANC_PQ_ENTID, ATIVO_NCOR_TOTAL, ATIVO_COR_INVENTARIOS, ATIVO_COR_CLIENTES, ATIVO_COR_ADIANTAMENTOS_FORNEC, ATIVO_COR_ESTADO_OUT_ENTES_PUB, ATIVO_COR_ACCIONISTAS_SOCIOS, ATIVO_COR_OUTRA_CONT_A_RECEBER, ATIVO_COR_DIFERIMENTOS, ATIVO_COR_ACT_FINAC_DETIDO_NEG, ATIVO_COR_OUTROS_ACT_FINANCEIR, ATIVO_COR_ACT_N_COR_DET_VENDA, ATIVO_COR_OUTROS_ACT_CORRENTES, ATIVO_COR_CAIXA_DEP_BANCARIOS, ATIVO_COR_TOTAL, ATIVO_TOTAL, CP_CAPITAL_REALIZADO, CP_ACCOES_PROPRIAS, CP_OUTROS_INSTR_CAP_PROPRIO, CP_PREMIOS_EMISSAO, CP_RESERVAS_LEGAIS, CP_OUTRAS_RESERVAS, CP_RESULTADOS_TRANSITADOS, CP_AJUST_EM_ACT_FINANCEIROS, CP_EXCENDENTES_REVALORIZACAO, CP_OUTRAS_VARIACAOES_CAP_PRO, CP_SOMA, CP_RESULTADO_LIQUIDO_PERIODO, CP_TOTAL, PASSIVO_NC_PROVISOES, PASSIVO_NC_FINANCIAMENTOS_OBTD, PASSIVO_NC_IMPOSTOS_DIFERIDOS, PASSIVO_NC_OUTRAS_CONTAS_A_PAGAR, PASSIVO_NC_TOTAL, PASSIVO_COR_FORNCEDORES, PASSIVO_COR_ADIANTA_DE_CLIENTES, PASSIVO_COR_ESTADO_OUT_ENT_PUB, PASSIVO_COR_ACCIONISTAS_SOCIOS, PASSIVO_COR_FINANCIAMENTOS_OBTD, PASSIVO_COR_OUT_CONTAS_A_PAGAR,

		PASSIVO_COR_DIFERIMENTOS, PASSIVO_COR_OUTROS_PASS_FINANCEI, PASSIVO_COR_OUTROS_PAS_CORRENTES, PASSIVO_COR_TOTAL, PASSIVO_TOTAL, CP_E_PASSIVO_TOTAL, 1_PESSOAL_NMP_PSERNR, 2_PESSOAL_NHT_PSERNR, 1_PESSOAL_NMP_PESS_REMUN_SE, 2_PESSOAL_NHT_PESS_REMUN_SE, 1_PESSOAL_NMP_PESS_N_REMUN, 2_PESSOAL_NHT_PESS_N_REMUN, 1_PESSOAL_NMP_PSE_TEMPO_COMPLETO, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, 1_PESSOAL_NMP_PSETC_REMUNERADAS, 2_PESSOAL_NHT_PSETC_REMUNERADAS, 1_PESSOAL_NMP_PSE_TEMPO_PARCIAL, 2_PESSOAL_NHT_PSE_TEMPO_PARCIAL, 1_PESSOAL_NMP_PSETP_REMUNERADAS, 2_N_PESSOAL_NHT_PSETP_REMUNERADAS, 1_PESSOAL_NMP_PSE_HOMENS, 2_PESSOAL_NHT_PSE_HOMENS, 1_PESSOAL_NMP_PSE_MULHERES, 2_PESSOAL_NHT_PSE_MULHERES, PESSOAL_NMP_PSE_INVEST_DESENV, 1_PESSOAL_NHT_PRESTAD_SERVICOS, 2_PESSOAL_NMP_PRESTAD_SERVICOS, PESSOAL_NHT_PESS_AGENC_TEMPOR, GASTOS_PESSOAL_TOTAL, GP_REMUN_ORGAOS_SOCIAIS, GP_ROS_PARTIC_LUCROS, GP_REMUN_PESSOAL, GP_RP_PARTIC_LUCROS, BENEF_POS_EMPREG, GP_BPE_PREMIOS_PENSOES, GP_BPE_OUTROS_BENEF, GP_INDEMINIZACOES, GP_ENCARG_REMUN, GP_SEG_ACID_TRAB_DOEN_PROF, GP_GASTOS_ACCAO_SOCIAL, OUTROS_GASTOS_PESSOAL, OGP_GASTOS_FORMACAO, OGP_GASTOS_FARDAMENTO, IAE_VENDAS, IAE_VENDAS_MERCADORIAS, IAE_VENDAS_PAISDRR, IAE_PREST_SERV, IAE_COMPRAS, IAE_FORN_SERV_EXTER, IAE_CMVMC, IAE_CMVMC_MERCADORIAS, IAE_CMVMC_MATER_PRIMAS, IAE_VAR_INVENT_PROD, IAE_NUM_MED_PESS_SERV, IAE_GASTOS_PESS, IAE_GASTOS_PESS_REMUN, IAE_GASTOS_PESS_OUTROS, IAE_AFT_QUANT_ESCR_LIQ_FIN, IAE_AFT_TOTAL_AQUIS, IAE_AFT_TOTAL_AQUIS_EDIF, IAE_AFT_ADIC_PERIOD_ACT, IAE_PI_QUANT_ESCR_LIQ_FIN, IAE_PI_TOTAL_AQUIS, IAE_PI_TOTAL_AQUIS_EDIF, 1_IMG_INT_VENDAS, 2_IMG_COM_VENDAS, 3_IMG_EC_VENDAS, 4_IMG_TOT_VENDAS, 1_IMG_INT_PREST_SERV, 2_IMG_COM_PREST_SERV, 3_IMG_EC_PREST_SERV, 4_IMG_TOT_PREST_SERV, 1_IMG_INT_COMPRAS, 2_IMG_COM_COMPRAS, 3_IMG_EC_COMPRAS, 4_IMG_TOT_COMPRAS, 1_IMG_INT_FORN_SEREXTERN, 2_IMG_COM_FORN_SEREXTERN, 3_IMG_EC_FORN_SEREXTERN, 4_IMG_TOT_FORN_SEREXTERN, 1_IMG_INT_AQUIS_ACT_FIX_TANG, 2_IMG_COM_AQUIS_ACT_FIX_TANG, 3_IMG_EC_AQUIS_ACT_FIX_TANG, 4_IMG_TOT_AQUIS_ACT_FIX_TANG, 1_IMG_INT_AQUIS_PROP_INVEST, 4_IMG_TOT_AQUIS_PROP_INVEST, 1_IMG_INT_AQUIS_ACT_INTANG, 2_IMG_COM_AQUIS_ACT_INTANG, 3_IMG_EC_AQUIS_ACT_INTANG, 4_IMG_TOT_AQUIS_ACT_INTANG, 1_IMG_INT_REND_SUPLEM, 2_IMG_COM_REND_SUPLEM,

		3_IMG_EC_REND_SUPLEM, 4_IMG_TOT_REND_SUPLEM, 1_IMG_INT_RS_SERSOCIAIS, 4_IMG_TOT_RS_SERSOCIAIS, 1_IMG_INT_RS_ALUG_EQUIP, 4_IMG_TOT_RS_ALUG_EQUIP, 1_IMG_INT_RS_EST_PROJ_ASS_TEC, 4_IMG_TOT_RS_EST_PROJ_ASS_TEC, 1_IMG_INT_RS_OUTROS, 2_IMG_COM_RS_OUTROS, 3_IMG_EC_RS_OUTROS, 4_IMG_TOT_RS_OUTROS, 1_IMG_INT_PM_VEND_PREST_SERV, 2_IMG_COM_PM_VEND_PREST_SERV, 3_IMG_EC_PM_VEND_PREST_SERV, 4_IMG_TOT_PM_VEND_PREST_SERV, 1_IMG_INT_PM_COMP_FORN_SER_EXT, 2_IMG_COM_PM_COMP_FORN_SER_EXT, 3_IMG_EC_PM_COMP_FORN_SER_EXT, 4_IMG_TOT_PM_COMP_FORN_SER_EXT, NIF_Prom_anon, Resumo/Distrito, Resumo/Concelho, Resumo/Cae, Resumo/Investimento, Resumo/Elegivel, Resumo/Dimensao, Resumo/Lst_Po, Resumo/Ot, Resumo/Pi, Resumo/Ti, Resumo/Nute_Norte, Resumo/Nute_Centro, Resumo/Nute_Lisboa, Resumo/Nute_Alentejo, Resumo/N_Pt_Pos, Parametros/Aut_Gestao, Parametros/Obj_Tema, Parametros/Prioridade, Parametros/Tipologia, Parametros/Centro, Promotor/Distrito, Promotor/Concelho, Promotor/Nat_Jur, Promotor/Cap_Social, Consultora/Nif_anon, Dimensao/Dimensao, Vantagenscomp/Estrategia, Dadosprojecto/N_Meses, Dadosprojecto/Inv_Geral, Dadosprojecto/Investimento, Dadosprojecto/Elegivel, Parametros/Norte, Analisemercados/Direcao, Critselb1/N_Mercados, Header/Informa_Db, Header/Proj_Aux, Parametros/Proj_Aux, Txtfinanc/Fonte, Incentivo/Elegivel, Incentivo/Dimensao, Incentivo/Pt_Qualif, Incentivo/Limite_Pt_Qualif, Incentivo/Aut_Gestao, Incentivo/Tx_Limite, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Ii, Incentivo/Base_Eleg, Incentivo/Base_Eleg_Bd, Incentivo/Tx_Major_Eleg, Incentivo/Tx_Major_Eleg_Bd, Incentivo/Valor_Major_Eleg, Incentivo/Valor_Major_Eleg_Bd, Incentivo/Incentivo, Incentivo/Incentivo_Nr, Incentivo/Incentivo_If, Descproj/Capacidade_Pre, Descproj/Capacidade_Pos, Inovtecn/Racio_Dr, EBITDA, EBIT, Total_Assets, Total_Liabilities, total_assets_to_total_liabilities, Working capital divided by total assets, Gross income divided by sales, total debt / total assets, earnings before tax and interest / total asset, Operating cash flow current liabilities, Accounts_Receivables_Turnover, Creditors_Turnover, Inventory_Turnover, Average_Collection_Period_For_Receivables, Average_Payment_Period_To_Creditors, Average_Turnover_Period_For_Inventories, Growth_Rate_Net_Sales_T1, Growth_Rate_Net_Sales_T2, Growth_Rate_Net_Sales_T3, Growth_Rate_Total_Assets_T1, Growth_Rate_Total_Assets_T2, Growth_Rate_Total_Assets_T3
manualFeat	22	EBITDA, EBIT, Total_Assets, Total_Liabilities, total_assets_to_total_liabilities, Working capital divided by total assets, Gross income divided by sales, total debt / total assets, earnings before tax and interest / total asset, Operating cash flow current liabilities, Accounts_Receivables_Turnover, Creditors_Turnover, Inventory_Turnover, Average_Collection_Period_For_Receivables,

		Average_Payment_Period_To_Creditors, Average_Turnover_Period_For_Inventories, Growth_Rate_Net_Sales_T1, Growth_Rate_Net_Sales_T2, Growth_Rate_Net_Sales_T3, Growth_Rate_Total_Assets_T1, Growth_Rate_Total_Assets_T2, Growth_Rate_Total_Assets_T3

Table 36. Outlier detection through z-score

Z-Score	
Rejected Points	Features
186	Resumo/Lst_Po, Parametros/Aut_Gestao
80	Growth_Rate_Total_Assets_T3
70	Resumo/Nute_Alentejo
66	Growth_Rate_Total_Assets_T2
65	ATIVO_TOTAL, CP_E_PASSIVO_TOTAL, Dadosprojecto/N_Meses, Total_Assets
64	PASSIVO_COR_ESTADO_OUT_ENT_PUB
63	CP_TOTAL
62	ATIVO_COR_CLIENTES
61	EBIT
59	EBITDA, Working capital divided by total assets
57	ATIVO_COR_TOTAL, PASSIVO_TOTAL, GP_ENCARG_REMUN, Total_Liabilities
56	RES_ANTES_DEPRECIACAO_GASTOS, RESULTADO_LIQUIDO_PERIODO, CP_RESULTADO_LIQUIDO_PERIODO, GP_REMUN_PESSOAL, Resumo/Nute_Lisboa
55	ATIVO_COR_INVENTARIOS, 2_PESSOAL_NHT_PSERNR, 2_PESSOAL_NHT_PESS_REMUN_SE, 2_PESSOAL_NHT_PSE_TEMPO_COMPLETO, 2_PESSOAL_NHT_PSETC_REMUNERADAS, Incentivo/Dimensao, Incentivo/Tx_Limite, Incentivo/Tx_Base, Incentivo/Tx_Major_I, Incentivo/Tx_Major_Eleg
54	GASTOS_DEPRECIACAO_AMORTIZA, PASSIVO_NC_FINANCIAMENTOS_OBTD, IAE_GASTOS_PESS_REMUN
53	CP_SOMA, 1_PESSOAL_NMP_PSERNR, 1_PESSOAL_NMP_PESS_REMUN_SE, 1_PESSOAL_NMP_PSE_TEMPO_COMPLETO, 1_PESSOAL_NMP_PSETC_REMUNERADAS, GP_REMUN_ORGAOS_SOCIAIS, IAE_NUM_MED_PESS_SERV, Resumo/Cae, Header/Informa_Db, total debt / total assets

52	VENDAS_SERVICOS_PRESTADOS, GASTOS_PESSOAL, GASTOS_PESSOAL_TOTAL, IAE_COMPRAS, IAE_GASTOS_PESS, 4_IMG_TOT_COMPRAS, Growth_Rate_Total_Assets_T1
51	ATIVO_COR_ESTADO_OUT_ENTES_PUB, IAE_VENDAS, IAE_VENDAS_PAISDRR, 4_IMG_TOT_VENDAS, Resumo/Investimento, Dadosprojecto/Investimento, Incentivo/Tx_Major_Ii
50	CMVMC, PASSIVO_COR_FORNCEDORES, PASSIVO_COR_TOTAL, IAE_CMVMC, Resumo/N_Pt_Pos, Promotor/Cap_Social, Dadosprojecto/Inv_Geral, Incentivo/Aut_Gestao
49	RESULTADO_OPERACIONAL, GP_SEG_ACID_TRAB_DOEN_PROF, Resumo/Elegivel, Dadosprojecto/Elegivel, Incentivo/Tx_Major_Eleg_Bd
48	CP_CAPITAL_REALIZADO, PASSIVO_COR_FINANCIAMENTOS_OBTD, 1_PESSOAL_NMP_PSE_HOMENS, 2_PESSOAL_NHT_PSE_HOMENS
47	RES_ANTES_IMPOSTOS, ATIVO_NCOR_TOTAL, IAE_CMVMC_MATER_PRIMAS, IAE_GASTOS_PESS_OUTROS
46	CP_OUTRAS_RESERVAS, 1_IMG_INT_VENDAS, 1_IMG_INT_COMPRAS
45	IAE_AFT_TOTAL_AQUIS, 4_IMG_TOT_AQUIS_ACT_FIX_TANG, Consultora/Nif_anon
44	IAE_AFT_TOTAL_AQUIS_EDIF, earnings before tax and interest / total asset
43	ATIVO_NCOR_FIXOS_TANGIVEIS, 2_PESSOAL_NHT_PSE_MULHERES, IAE_AFT_QUANT_ESCR_LIQ_FIN
42	ATIVO_COR_OUTRA_CONT_A_RECEBER, Incentivo/Limite_Pt_Qualif
41	2_IMG_COM_VENDAS
40	IMPOSTO_RENDIMENTO_PERIODO, 1_PESSOAL_NMP_PSE_MULHERES
39	SUBSIDIOS_EXPLORACAO, JUROS_GAST_SIMILARES_SUPORT, CP_OUTRAS_VARIACAOES_CAP_PRO, PASSIVO_NC_TOTAL, PASSIVO_COR_OUT_CONTAS_A_PAGAR, OGP_GASTOS_FORMACAO
38	3_IMG_EC_VENDAS, 2_IMG_COM_FORN_SEREXTERN, 3_IMG_EC_FORN_SEREXTERN
37	ATIVO_COR_CAIXA_DEP_BANCARIOS, 1_PESSOAL_NMP_PSE_TEMPO_PARCIAL, 2_IMG_COM_AQUIS_ACT_FIX_TANG
36	OUTROS_GASTOS_PERDAS, 2_IMG_COM_COMPRAS, 1_IMG_INT_AQUIS_ACT_FIX_TANG
35	2_PESSOAL_NHT_PSE_TEMPO_PARCIAL
34	CP_RESULTADOS_TRANSITADOS, 1_PESSOAL_NMP_PSETP_REMUNERADAS, IAE_VENDAS_MERCADORIAS, Incentivo/Pt_Qualif

33	ATIVO_NCOR_PART_FINAN_EQV_PAT, PASSIVO_COR_ADIANTA_DE_CLIENTES, 2_IMG_COM_PREST_SERV, Resumo/Ot, Resumo/Pi, Resumo/Ti, Parametros/Obj_Tema, Parametros/Prioridade, Parametros/Tipologia
32	2_N_PESSOAL_NHT_PSETP_REMUNERADAS, 3_IMG_EC_COMPRAS
31	FSE, OUTROS_REDIMENTOS_GANHOS, IAE_FORN_SERV_EXTER, 4_IMG_TOT_FORN_SEREXTERN
30	ATIVO_NCOR_IMPOSTOS_DIFERIDOS, IAE_CMVMC_MERCADORIAS, 1_IMG_INT_FORN_SEREXTERN
28	PASSIVO_COR_DIFERIMENTOS, 1_PESSOAL_NMP_PESS_N_REMUN, PESSOAL_NMP_PSE_INVEST_DESENV, 1_PESSOAL_NHT_PRESTAD_SERVICOS, OGP_GASTOS_FARDAMENTO, Resumo/Dimensao, Dimensao/Dimensao
27	IAE_PREST_SERV, 4_IMG_TOT_PREST_SERV
26	2_IMG_COM_REND_SUPLEM, 2_IMG_COM_RS_OUTROS
25	JUROS_RENDIME_SIMILARES_OBT, ATIVO_COR_OUTROS_ACT_FINANCEIR, 4_IMG_TOT_REND_SUPLEM
24	ATIVO_NCOR_OUTROS_ACT_FINANC, 1_IMG_INT_PREST_SERV
23	TRABALHOS_PROPRIA_ENTIDADE, ATIVO_NCOR_INTANGIVEIS, CP_AJUST_EM_ACT_FINANCEIROS, CP_EXCENDENTES_REVALORIZACAO, GP_GASTOS_ACCAO_SOCIAL, IAE_AFT_ADIC_PERIOD_ACT
22	VARIA_INVENTARIOS_PRODUCAO, PASSIVO_NC_PROVISOES, GP_INDEMINIZACOES, IAE_VAR_INVENT_PROD, 4_IMG_TOT_RS_OUTROS, 4_IMG_TOT_PM_COMP_FORN_SER_EXT
21	ATIVO_NCOR_PROPRI_INVESTIMENTO, IAE_PI_QUANT_ESCR_LIQ_FIN, Incentivo/Elegivel, Incentivo/Incentivo, Incentivo/Incentivo_Nr
20	IMPARIDADE_INVENTARIOS, ATIVO_COR_ACCIONISTAS_SOCIOS, CP_ACCOES_PROPRIAS, GP_RP_PARTIC_LUCROS
19	IMPARIDADE_DIVIDAS_A_RECEBER, PASSIVO_NC_IMPOSTOS_DIFERIDOS, 1_IMG_INT_PM_COMP_FORN_SER_EXT, Inovtecn/Racio_Dr
18	ATIVO_COR_ACT_FINAC_DETIDO_NEG, PASSIVO_COR_ACCIONISTAS_SOCIOS, 2_PESSOAL_NMP_PRESTAD_SERVICOS, GP_ROS_PARTIC_LUCROS, Txtfinanc/Fonte, Incentivo/Valor_Major_Eleg
17	CP_RESERVAS_LEGAIS, 3_IMG_EC_AQUIS_ACT_FIX_TANG, 1_IMG_INT_REND_SUPLEM, Incentivo/Base_Eleg

16	BENEF_POS_EMPREG, OUTROS_GASTOS_PESSOAL, 4_IMG_TOT_AQUIS_ACT_INTANG, Inventory_Turnover
15	GANHOS_PERDAS_SUBSIDIARIAS, PROVISOES, ATIVO_NCOR_ACCIONISTAS_SOCIOS, ATIVO_COR_ADIANTAMENTOS_FORNEC, 3_IMG_EC_PREST_SERV, 1_IMG_INT_AQUIS_ACT_INTANG, 3_IMG_EC_PM_VEND_PREST_SERV
14	ATIVO_COR_OUTROS_ACT_CORRENTES, PASSIVO_COR_OUTROS_PAS_CORRENTES, PESSOAL_NHT_PESS_AGENC_TEMPOR, 2_IMG_COM_PM_COMP_FORN_SER_EXT, Incentivo/Base_Eleg_Bd, Incentivo/Valor_Major_Eleg_Bd
13	2_PESSOAL_NHT_PESS_N_REMUN, 1_IMG_INT_RS_OUTROS, Incentivo/Incentivo_If
12	CP_PREMIOS_EMISSAO, Header/Proj_Aux, Parametros/Proj_Aux
11	PASSIVO_NC_OUTRAS_CONTAS_A_PAGAR
10	IAE_PI_TOTAL_AQUIS, 1_IMG_INT_AQUIS_PROP_INVEST, 4_IMG_TOT_AQUIS_PROP_INVEST, 3_IMG_EC_RS_OUTROS
9	IMPARIDADE_INVENTA_N_AMORT_INVESTIMENTOS_N_DEPRECIAVEIS, ATIVO_NCOR_INV_FINANC_PQ_ENTID, CP_OUTROS_INSTR_CAP_PROPRIO, GP_BPE_PREMIOS_PENSOES, Creditors_Turnover, Average_Collection_Period_For_Receivables
8	IAE_PI_TOTAL_AQUIS_EDIF, 2_IMG_COM_AQUIS_ACT_INTANG, 3_IMG_EC_AQUIS_ACT_INTANG, 4_IMG_TOT_RS_ALUG_EQUIP, 2_IMG_COM_PM_VEND_PREST_SERV, Growth_Rate_Net_Sales_T1
7	AUMENTOS_RED_JUSTO_VALOR, ATIVO_COR_ACT_N_COR_DET_VENDA, GP_BPE_OUTROS_BENEF, 1_IMG_INT_RS_SERSOCIAIS, 4_IMG_TOT_RS_SERSOCIAIS, 1_IMG_INT_RS_ALUG_EQUIP, Descproj/Capacidade_Pos, Operating cash flow current liabilities
6	ATIVO_NCOR_PART_FIN_OUTROS_MET, ATIVO_COR_DIFERIMENTOS, PASSIVO_COR_OUTROS_PASS_FINANCEI, 4_IMG_TOT_RS_EST_PROJ_ASS_TEC, Descproj/Capacidade_Pre
5	OUTRAS_IMPARIDADES, ATIVO_NCOR_GOODWILL, 3_IMG_EC_REND_SUPLEM, 1_IMG_INT_RS_EST_PROJ_ASS_TEC, Vantagenscomp/Estrategia, Growth_Rate_Net_Sales_T2
4	3_IMG_EC_PM_COMP_FORN_SER_EXT, Gross income divided by sales, Average_Turnover_Period_For_Inventories
3	Promotor/Nat_Jur, total_assets_to_total_liabilities, Growth_Rate_Net_Sales_T3

1	1_IMG_INT_PM_VEND_PREST_SERV, 4_IMG_TOT_PM_VEND_PREST_SERV, Accounts_Receivables_Turnover, Average_Payment_Period_To_Creditors

Table 37. Parameters used in models training

Model	Parameters
LogisticRegression	penalty='l2' dual=False tol= 0.0001 C=1.0 fit_intercept=True intercept_scaling=1 class_weight=None solver='lbfgs' max_iter=100 multi_class='auto' verbose=0 warm_start=False n_jobs=None l1_ratio=None
SupportVectorClassification	C=1.0 kernel='rbf' degree=3 gamma='scale' coef0=0.0 shrinking=True probability=False tol=0.001 cache_size=200 class_weight=None verbose=False max_iter=-1 decision_function_shape='ovr' break_ties=False
LinearSupportVectorClassification	penalty='l2'

	loss='squared_hinge' dual=True tol=0.0001 C=1.0 multi_class='ovr' fit_intercept=True intercept_scaling=1 class_weight=None verbose=0 max_iter=1000
K-NearestNeighbors	n_neighbors=5 weights='uniform' algorithm='auto' leaf_size=30 p=2 metric='minkowski' metric_params=None n_jobs=None
GaussianNaiveBayes	var_smoothing=0.000000001
Perceptron	penalty=None alpha=0.0001 fit_intercept=True max_iter=1000 tol=1e-3 shuffle=True verbose=0 eta0=1.0 n_jobs=None early_stopping=False validation_fraction=0.1 n_iter_no_change=5 class_weight=None warm_start=False
DecisionTree	criterion='gini'

	splitter='best' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None max_leaf_nodes=None min_impurity_decrease=0.0 min_impurity_split=None class_weight=None
RandomForest	n_estimators=100 criterion='gini' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features='auto' max_leaf_nodes=None min_impurity_decrease=0.0 min_impurity_split=None bootstrap=True oob_score=False n_jobs=None verbose=0 warm_start=False class_weight=None
MLP Classifier	hidden_layer_sizes=(100,) activation='relu' solver='adam' alpha=0.0001 batch_size='auto' learning_rate='constant' learning_rate_init=0.001 power_t=0.5

	max_iter=200 shuffle=True tol=1e-4 verbose=False warm_start=False momentum=0.9 nesterovs_momentum=True early_stopping=False validation_fraction=0.1 beta_1=0.9 beta_2=0.999 epsilon=1e-8 n_iter_no_change=10 max_fun=15000
XGBoost	base_score=0.5 booster='gbtree' colsample_bylevel=1 colsample_bynode=1 colsample_bytree=1 gamma=0 gpu_id=-1 importance_type='gain' interaction_constraints='' learning_rate=0.3 max_delta_step=0 max_depth=6 min_child_weight=1 missing=None monotone_constraints='()' n_estimators=100 n_jobs=8 num_parallel_tree=1 objective='binary:logistic' random_state=0

	reg_alpha=0 reg_lambda=1 scale_pos_weight=1 subsample=1 tree_method='exact' validate_parameters=1

Table 38. Constants features removed from dataset

Column
resultado_das_atividades_descontinuadas
IAE_CMVMC_ACT_BIOL
Uploads/Aplicavel_12
3_IMG_EC_RS_SERSOCIAIS
2_IMG_COM_RS_SERSOCIAIS
3_IMG_EC_AQUIS_PROP_INVEST
2_IMG_COM_AQUIS_PROP_INVEST
Header/Exportado
Dimensao_Empresa
2_IMG_COM_RS_ROYALTIES
GP_BPE_PLAN_CONTRIB_OUTROS
GP_BPE_PLAN_CONTRIB_ORG_SOC
Header/Reg_Pas
Resumo/Internacional
CP_DIVIDENDOS_ANTECIPADOS
Declaracoes/Geral_2
Uploads/Upload_12
Resumo/Turismo
2_IMG_COM_RS_EST_PROJ_ASS_TEC
PASSIVO_COR_PASS_FINANC_DET_NEG
PASSIVO_COR_PAS_NC_DETIDOS_VENDA
3_IMG_EC_RS_EST_PROJ_ASS_TEC
ATIVO_NCOR_BIOLOGICOS
ATIVO_COR_ACTIVOS_BIOLOGICOS
PASSIVO_NC_RESP_BENEF_POS_EMPREG
1_IMG_INT_RS_ROYALTIES
3_IMG_EC_RS_ALUG_EQUIP
Uploads/Upload_08
2_IMG_COM_RS_ALUG_EQUIP
Uploads/Aplicavel_08
Checklist/Igualdade_5
Checklist/Igualdade_2
Declaracoes/Eleg_Prom_5
Declaracoes/Obrig_1
Declaracoes/Eleg_Proj_1
Declaracoes/Eleg_Prom_6

Declaracoes/Eleg_Prom_3
Declaracoes/Eleg_Prom_4
Declaracoes/Eleg_Prom_1
IAE_PI_ADIC_PERIO_PROP_INV
3_IMG_EC_RS_ROYALTIES
Declaracoes/Eleg_Prom_2
Checklist/Igualdade_3
IAE_VENDAS_ACT_BIOL
Header/Extensao
Checklist/Integracao_3
Checklist/Integracao_2
Checklist/Integracao_1
Checklist/Conciliacao_2
Dadosprojecto/Inv_Diaspora
Checklist/Aval_Global_2
IMPARIDADE_INVEST_AMORTIZ_DEPRECIAVEIS
Checklist/Conciliacao_1
Checklist/Igualdade_1
Checklist/Aval_Global_1
Provere/Pergunta
Checklist/Prevencao_1
Checklist/Prevencao_2
Checklist/Igualdade_4
Declaracoes/Geral_4
4_IMG_TOT_RS_ROYALTIES
Promotor/Fins_Lucro
Resumo/Nute_Algarve
Parametros/Algarve
Incentivo/Tx_Major_Iv
Incentivo/Cap_Proprios
Incentivo/Dispensa_Ic
Paramproj/Param_1
Resumo/Icep_75
Impactoemp/Impacto
N_Proj_anon
CAE_SUBCLASSE
DATA_RECEPCAO
NIF_anon
ANO_EXERCICIO

Table 39. Description of features used in best experiences

Financial ratio (feature name)	**Description**
ATIVO_COR_ACCIONISTAS_SOCIOS	Current Assets Shareholders and Partners
Incentivo/Tx_Major_Ii	Incentive/Major Rate
Promotor/Concelho	Promoter/Council
ATIVO_COR_ESTADO_OUT_ENTES_PUB	Current Assets State and Other Public Entities
ATIVO_COR_DIFERIMENTOS	Current Assets Deferred Charges
ATIVO_COR_OUTRA_CONT_A_RECEBER	Current Assets Other Receivables

CP_RESERVAS_LEGAIS	Legal Reserves
Resumo/N_Pt_Pos	Summary/Number of Positions
Growth_Rate_Net_Sales_T3	Growth Rate of Net Sales (3 years before application)
Resumo/Elegivel	Summary/Eligible
3_IMG_EC_REND_SUPLEM	Image EC Supplementary Income
Txtfinanc/Fonte	Financial Text/Source
1_PESSOAL_NMP_PSE_MULHERES	Average number of renumerated women
Incentivo/Tx_Major_Eleg	Incentive/Major Rate Eligible
IAE_CMVMC_MERCADORIAS	IAE CMVMC Merchandise
CP_OUTRAS_RESERVAS	Other Equity Reserves
Growth_Rate_Total_Assets_T1	Growth Rate of Total Assets (3 years before application)
Incentivo/Tx_Major_I	Incentive/Major Rate
IAE_CMVMC_MATER_PRIMAS	CMVMC Raw Materials
Promotor/Distrito	Promoter/District
Resumo/Investimento	Summary/Investment
2_PESSOAL_NHT_PSETC_REMUNERADAS	Total number of hours of paid staff
2_PESSOAL_NHT_PSERNR	Total number of hours of non-paid staff
CP_E_PASSIVO_TOTAL	Equity and Total Liabilities
Resumo/Cae	Summary/CAE (Economic Activity Code)
Incentivo/Pt_Qualif	Incentive/Qualification Points
GP_SEG_ACID_TRAB_DOEN_PROF	Group Insurance for Work Accidents and Occupational Diseases
OUTROS_REDIMENTOS_GANHOS	Other Income and Gains
GASTOS_PESSOAL	Personnel Expenses
Dadosprojecto/N_Meses	Project Data/Number of Months
Incentivo/Dimensao	Incentive/Size
Total_Assets	Total Assets
Growth_Rate_Net_Sales_T2	Growth Rate of Net Sales (2 years before application)
1_PESSOAL_NMP_PSETC_REMUNERADAS	Average number of paid staff
Grossincomedividedbysales	Gross Income Divided by Sales
Incentivo/Incentivo_Nr	Incentive/Incentive Number
PASSIVO_COR_OUT_CONTAS_A_PAGAR	Current Liabilities Other Accounts Payable
ATIVO_NCOR_INV_FINANC_PQ_ENTID	Non-Current Assets Investment in Financial Assets of Public Interest
NIF_Prom_anon	Promoter's Tax Identification Number (NIF)
IAE_PREST_SERV	IAE Provision of Services
IAE_VAR_INVENT_PROD	IAE Variation in Inventory of Products
SUBSIDIOS_EXPLORACAO	Subsidies for Operations
IAE_VENDAS_PAISDRR	IAE Sales to Countries with Double Taxation Relief
Resumo/Nute_Norte	Summary/Nute North
2_PESSOAL_NHT_PSE_MULHERES	Total number of hours women
IAE_COMPRAS	IAE Purchases
1_IMG_INT_FORN_SEREXTERN	Image International Suppliers and External Services
Incentivo/Tx_Limite	Incentive/Limit Rate

IAE_AFT_QUANT_ESCR_LIQ_FIN	Quantity of Tangible Fixed Assets in the Financial Liquidation
IMPOSTO_RENDIMENTO_PERIODO	Income Tax for the Period
ATIVO_COR_INVENTARIOS	Current Assets Inventories
PASSIVO_TOTAL	Total Liabilities
CP_RESULTADO_LIQUIDO_PERIODO	Current Liabilities Net Result for the Period
2_IMG_COM_VENDAS	Image Commercial Sales
Inventory_Turnover	Inventory Turnover
GASTOS_PESSOAL_TOTAL	Total Personnel Expenses
total_assets_to_total_liabilities	Total Assets to Total Liabilities Ratio
IAE_AFT_TOTAL_AQUIS_EDIF	Total Tangible Fixed Assets Acquisition Ratio
Resumo/Dimensao	Summary/Size
Incentivo/Tx_Base	Incentive/Base Rate
1_IMG_INT_AQUIS_ACT_FIX_TANG	Image Acquisition of Tangible Fixed Assets
IAE_VENDAS_MERCADORIAS	IAE Sales of Goods
3_IMG_EC_COMPRAS	Image Purchases
RES_ANTES_DEPRECIACAO_GASTOS	Income before Depreciation and Expenses
ATIVO_NCOR_TOTAL	Non-Current Assets Total
totaldebt/totalassets	Total Debt to Total Assets Ratio
1_IMG_INT_REND_SUPLEM	Image Supplementary Income
GP_ENCARG_REMUN	Employee Expense Ratio
PASSIVO_NC_FINANCIAMENTOS_OBTD	Non-Current Liabilities Borrowings Obtained
ATIVO_NCOR_PART_FINAN_EQV_PAT	Non-Current Assets Equity and Other Financial Instruments
RES_ANTES_IMPOSTOS	Income before Taxes
CP_TOTAL	Total Equity
PASSIVO_COR_ESTADO_OUT_ENT_PUB	Current Liabilities State and Other Public Entities
VENDAS_SERVICOS_PRESTADOS	Sales of Services Provided
Resumo/Concelho	Summary/Council
Incentivo/Incentivo	Incentive
PASSIVO_COR_OUTROS_PAS_CORRENTES	Current Liabilities Other Current Liabilities
CP_OUTRAS_VARIACAOES_CAP_PRO	Other Changes in Capital and Reserves
Operatingcashflowcurrentliabilities	Operating Cash Flow to Current Liabilities Ratio
earningsbeforetaxandinterest/totalasset	Earnings before Tax and Interest to Total Assets Ratio
GP_REMUN_ORGAOS_SOCIAIS	Executive Compensation Personnel Expenses
2_PESSOAL_NHT_PESS_REMUN_SE	Total number of hours of paid staff
PASSIVO_COR_FORNCEDORES	Current Liabilities Suppliers
IAE_GASTOS_PESS	IAE Personnel Expenses
GASTOS_DEPRECIACAO_AMORTIZA	Depreciation and Amortization Expenses
Incentivo/Tx_Major_Eleg_Bd	Incentive/Major Rate Eligibility Bond
2_PESSOAL_NHT_PSE_TEMPO_COMPLETO	Total number of hours of full-time staff
Resumo/Distrito	Summary/District
Incentivo/Aut_Gestao	Incentive/Management Authority
4_IMG_TOT_VENDAS	Image Total Sales
Incentivo/Base_Eleg	Incentive/Eligibility Base